Handbook for Assessment in the Service of Learning

Volume III

Examples of Assessment in the Service of Learning

Edited by
Eric M. Tucker
Howard T. Everson
Eva L. Baker
Edmund W. Gordon

Handbook for Assessment in the Service of Learning

Volume III

Examples of Assessment
in the Service of Learning

UMassAmherst
University Libraries

Edited by
Eric M. Tucker
Howard T. Everson
Eva L. Baker
Edmund W. Gordon

Series Editors

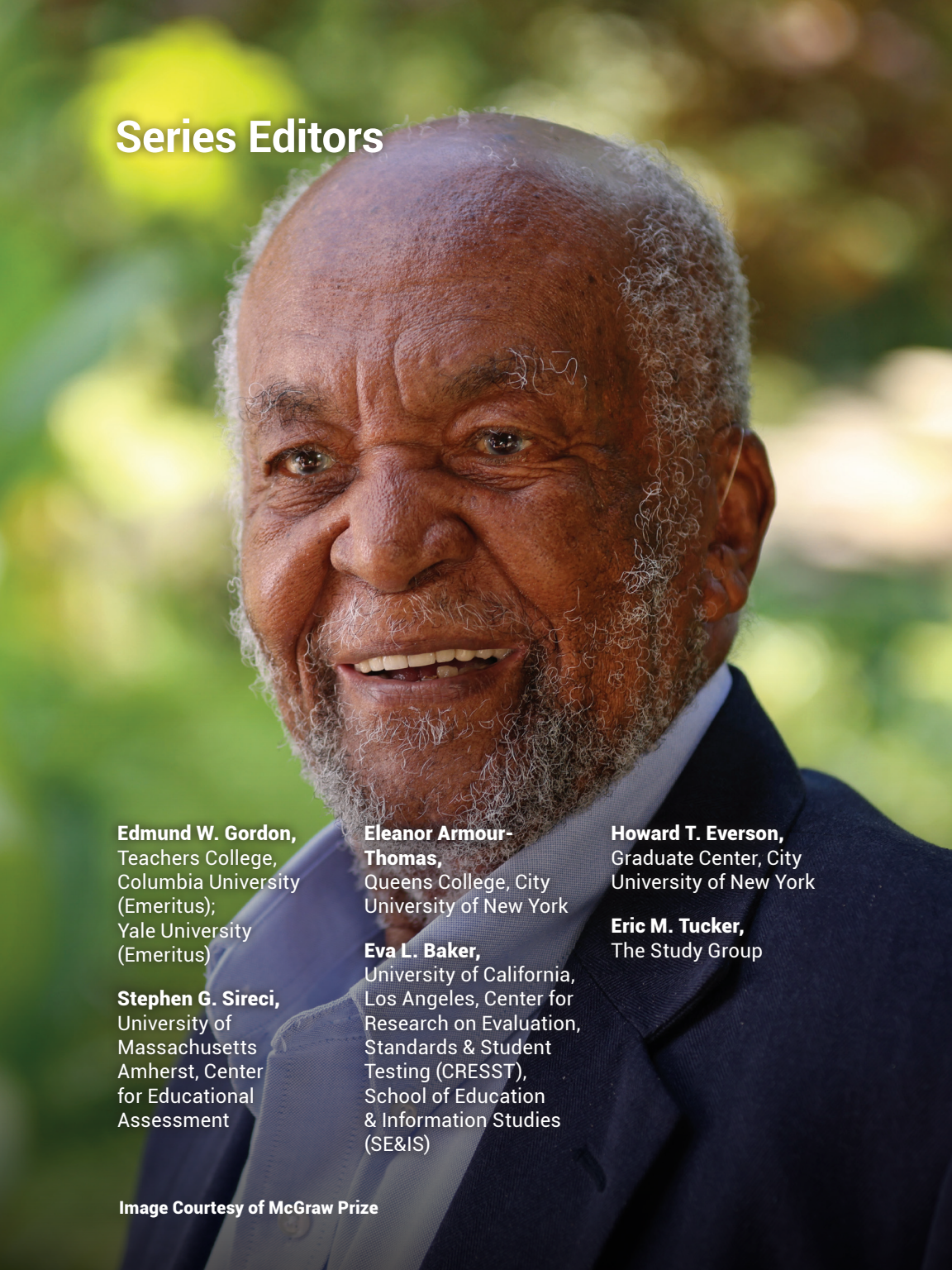


Image Courtesy of McGraw Prize

UMassAmherst

University Libraries

Handbook for Assessment in the Service of Learning, Volume III: Examples of Assessment in the Service of Learning ©

First edition published October 2025
by the University of Massachusetts Amherst
https://openpublishing.library.umass.edu/

DOI: 10.7275/s78z-y897
ISBN: 978-1-945764-35-6

Cover Design by Dezudio
Book Design by The Study Group

The suggested citation for this handbook is: Tucker, E. M., Everson, H. T., Baker, E. L., & Gordon, E. W. (Eds.). (2025). *Handbook for Assessment in the Service of Learning, Volume III: Examples of assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Contents

Volume Contributors

Eleanor Armour-Thomas, Queens College, City University of New York (Emeritus)

Eva L. Baker, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Hee Jin Bang, Age of Learning

Anastasia Betts, Learnology Labs

Jack Buckley, Roblox

Jennifer Charlot, RevX

Gregory K. W. K. Chung, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST)

Kimberly Cockrell, The Achievement Network, Ltd.

Kelly Corrado, PBS KIDS

Girlie C. Delacruz, Northeastern University

Kristen DiCerbo, Khan Academy

Clarissa Deverel-Rico, BSCS Science Learning

David S. Escoffery, Educational Testing Service

Howard T. Everson, Graduate Center, City University of New York

Cosimo Felline, PBS KIDS

Edmund W. Gordon, Teachers College, Columbia University (Emeritus); Yale University (Emeritus)

Sunil Gunderia, Age of Learning

Emily C. Hanno, MDRC

JoAnn Hsueh, MDRC

Diana Hughes, Age of Learning

Paul G. LeMahieu, Carnegie Foundation for the Advancement of Teaching; University of Hawai'i, Mānoa

Lei Liu, Educational Testing Service

Ou Lydia Liu, Educational Testing Service

Silvia Lovato, PBS KIDS

Susan Lyons, Lyons Assessment Consulting

Elizabeth Mokyr Horner, Gates Foundation

Michelle Odemwingie, The Achievement Network, Ltd.

Saskia Op den Bosch, RevX

V. Elizabeth Owen, Age of Learning

Trevor Packer, College Board

James W. Pellegrino, University of Illinois Chicago

Ximena A. Portilla, MDRC

Jeremy D. Roberts, PBS KIDS

Mary-Celeste Schreuder, The Achievement Network, Ltd.

David Sherer, Carnegie Foundation for the Advancement of Teaching

Stephen G. Sireci, University of Massachusetts Amherst, Center for Educational Assessment

Erica Snow, Roblox

Rebecca Stone-Danahy, College Board

Rebecca Sutherland, Reading Reimagined/AERDF

Natalya Tabony, College Board

Carrie Townley-Flores, Rapid Online Assessment of Reading (ROAR) at Stanford University

Eric M. Tucker, The Study Group

Jessica Wise Younger, PBS KIDS

Constance Yowell, Northeastern University

Credits

We gratefully acknowledge the leadership and dedication of the editorial team, whose vision, commitment, and expertise made the Handbook for Assessment in the Service of Learning series possible.

Series Editors
Edmund W. Gordon
Stephen G. Sireci
Eleanor Armour-Thomas
Eva L. Baker
Howard T. Everson
Eric M. Tucker

Managing Editors
Eric M. Tucker
Sheryl L. Gómez

We owe a profound debt of gratitude to *Professor Edmund W. Gordon* for his visionary conceptual leadership, which provided the inspiration and foundation for this Series. His friendship and decades-long commitment to scholarship that advances understanding of assessment in the service of learning has been the fountainhead throughout this project.

We gratefully acknowledge the *Gordon Seminar for Assessment in the Service of Learning,* housed at the Edmund W. Gordon Institute for Advanced Study at Teachers College, for its pivotal role in supporting the initial conceptualization of this Handbook. Convened by Professor Gordon starting in 2020 to advance the charge of the Gordon Commission for the Future of Assessment in Education, the Seminar provided a critical forum in which many of the ideas in these volumes were presented, debated, and refined. For over fifty years, the Gordon Institute has used advocacy, demonstration, evaluation, information dissemination, research, and technical assistance to study and seek to improve the quality of life chances of communities of color through education in urban contexts.

We acknowledge *The Study Group* for stewarding the project to publication, including by assuming the project lead and managing editorial functions. The Study Group coordinated the solicitation and review of chapters, managed author communications, oversaw the copyediting, layout, and design, and delivered the manuscripts to the publisher. This leadership was essential to the Handbook's successful completion.

Acknowledgements

The *Handbook for Assessment in the Service of Learning* is the product of a dedicated community of scholars and practitioners, but we owe our most profound debt of gratitude to Professor Edmund W. Gordon. His scholarship provides the foundational inspiration and ethical compass for this series. From inception, Professor Gordon contributed the precious heirloom seed concepts planted and cultivated into the Handbook chapters. As convener of the Gordon Seminar for Assessment in the Service of Learning, housed at Teachers College, Columbia University, he fostered the rigorous inquiry and in-depth discussions that strengthened the core ideas forming the intellectual bedrock of these volumes. He challenged us to be ambitious, and his guidance was the essential element that sustained this collaboration. This Handbook series would not exist without him, and we are honored to carry forward his legacy.

We extend our sincere thanks to the Series Editors—Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker—whose collective vision, expertise, and commitment were instrumental in shaping the intellectual direction of this series. The Volume Co-Editors were fundamental in securing the quality of the scholarship within each volume. Guiding the operational and logistical dimensions of this complex process were our Managing Editors, Eric M. Tucker and Sheryl L. Gómez, who earned our thanks for their remarkable efforts in steering the processes from start to finish.

The conceptual origins of this Handbook series are rooted in the seminal work of the *Gordon Commission on the Future of Assessment in Education* (2011–2013). The Commission planted seeds that are beginning to come to fruition in these volumes. The Edmund W. Gordon Institute for Advanced Study in Education at Teachers College, Columbia University, now embracing its fifth decade, has served as the vital intellectual home for Professor Gordon and the ambitious projects he undertakes. We thank the Gordon Institute for the support that makes Professor Gordon's prolific scholarly life possible during his 104th year. We are grateful to Ezekiel Dixon-Román, the Gordon Institute's Director, and Paola Heincke, who have been steadfast partners for Professor Gordon's vision. We extend special thanks to

Jonthon Coulson. His intellect, writing, curiosity, sense of adventure, and kindness left an indelible mark on the Seminar and this Handbook, and we are particularly grateful for his stewardship and foundational organizational and conceptual contributions during the program's formative iterations.

The Gordon Seminar for Assessment in the Service of Learning formed a community of inquiry. The thoughtful feedback from its participants provided the intellectual space to test, develop, and strengthen the core ideas in these volumes. We offer profound appreciation to the Seminar's core participants: Eleanor Armour-Thomas, Aneesha Badrinarayan, Eva L. Baker, Randy Bennett, Susan M. Brookhart, Greg Chung, Madhabi Chatterji, Jonthon Coulson, Linda Darling-Hammond, Ezekiel J. Dixon-Román, Richard Durán, Howard T. Everson, Sheryl L. Gómez, Edmund W. Gordon, Kris D. Gutiérrez, Kenji Hakuta, Gerunda B. Hughes, Neal Kingston, Carol D. Lee, John Lee, Paul G. LeMahieu, Pamela Moss, Temple Lovelace, Susan Lyons, Robert J. Mislevy, Maria Elena Oliveri, Roy Pea, Jennifer Randall, Stephen G. Sireci, Eric M. Tucker, Ernest Washington.

We also thank the many distinguished colleagues who, as Seminar presenters and guests, challenged our assumptions and enriched our dialogue with their cutting-edge research: Itzel Aceves, Ryan Baker, Yoav Bergner, Abby Benedetto, John Behrens, Lauren Bierbaum, Jill Burstein, Tony Bryk, Pamela Cantor, Andy Calkins, Auditi Chakravarty, Andrew Dalton, Jacqueline Darvin, Kristen DiCerbo, Fabienne Doucet, Kadriye Ercikan, Dave Escoffery, Tianying (Teanna) Feng, Natalie Foster, Peter Gault, Jim Gee, E. Wyatt Gordon, Sunil Gunderia, Khaled J. Ismail, Fiona Hinds, Kristen Huff, JoAnn Hsueh, Neil T. Heffernan, Elizabeth Mokyr Horner, Rebecca Kockler, Timothy Knowles, Michael Kearns, Jade Caines, Richard Lerner, Lydia Liu, Maxine McKinney de Royston, Orrin Murray, Jasmine McBeath Nation, Britt Neuhaus, Osarugue Michelle Odemwingie, Andreas Oranje, Trevor Packer, Luciana Parisi, James W. Pellegrino, William Penuel, Mario Piacentini, Ramona Pierson, Elizabeth Redman, Jeremy Roberts, Barbara Rogoff, Maheen Sahoo, Amit Sevak, Lorrie A. Shepard, Laura Slover, Jim Shelton, Valerie Shute, Kim Smith, Rebecca Stone-Danahy, Natalya Tabony, Sylvane Vaccarino, Arthur VanderVeen, Alina von Davier, Alyssa Wise, and Jason Yeatman. We are grateful to all who lent their expertise to this collaborative process, and we offer special thanks to Eleanor Armour-Thomas, Eric Tucker, and Sheryl Gómez for expertly moderating and organizing the Seminar. These sessions provided vital feedback on the Handbook chapters and framing as a work-in-progress.

We are thankful to the colleagues who participated in the AERA Honorary Presidential Sessions during annual meetings of the American Educational Research Association, providing a crucial platform for engaging with a range of viewpoints. The participants included: Brenda Allen, C. Malik Boykin, M. C. Brown, E. Wyatt Gordon, Jessica Heppen, Gabriela Lopez, Jamie Olson McKee, James L. Moore III, Na'ilah Suad Nasir, Anne Marie Núñez, Roberto J. Rodríguez, Timothy E. Sams, Mark Schneider, Matthew Soldner, LaVerne Evans Srinivasan, Claude Steele, Erica N. Walker, Amy Stuart Wells, Lester W. Young, Jr., and Elham Zandvakili.

Furthermore, a series of academic sessions convened to honor Professor Gordon's 100th birthday and his extraordinary legacy proved essential to this project's development. We extend our sincere gratitude to the host institutions, including Teachers College, Columbia University; University of California, Los Angeles; University of California, Santa Barbara; University of Massachusetts Amherst; and University of Texas at Austin. We thank the organizers and participants of these conferences; their engagement helped sharpen this Handbook series.

At the heart of this project are the contributions of the nearly 90 chapter authors whose collective scholarship forms the core of the Handbook. We thank them for their expertise and commitment. We are profoundly grateful to the series and volume editors, and peer reviewers, whose insightful critiques strengthened the quality, clarity, and coherence of each chapter.

We acknowledge The Study Group for its indispensable role in stewarding this project from conception to publication. Eric M. Tucker, Sheryl L. Gómez, Lauren Cutuli, Ciara Scott, and their team, expertly coordinated the complex processes of author communication, manuscript preparation and review, and production with skill and dedication. We are grateful to the University of Massachusetts Amherst Libraries for their partnership and commitment to open-access scholarship. The design and production of the Study Group and Dezudio design teams transformed our manuscripts into a polished and accessible final publication. This includes Ian Boly, Melissa Neely, Klaus Bellon, Ashley Deal, and Raelynn O'Leary.

We dedicate this Handbook to the memory of our cherished colleagues from the Gordon Commission who passed away during this work: Jamal Abedi, Lloyd Bond, A. Wade Boykin, Carl F. Kaestle, James Greeno, Stafford Hood, Robert J. Mislevy, and Lee Shulman. Their wisdom, friendship, and spirit were foundational

to this project, and their loss is deeply felt. We also remember all others from our community who have passed on; their contributions are woven into the fabric of this work, and we honor them with gratitude and respect.

Finally, on a personal note, we thank our families and friends for their support and patience throughout this journey. Our loved ones' understanding and encouragement sustained us through the long hours of research, writing, and editing. Each of the editors is grateful to those mentors and colleagues who offered personal support and guidance along the way—while too numerous to name here, please know that your influence has been invaluable.

In closing, we view the *Handbook for Assessment in the Service of Learning* as the harvest of many years of collaborative effort—a harvest that we are delighted to share with the world. Professor Gordon used an agricultural metaphor to describe this project, speaking of selecting and sowing conceptual seeds, cultivating fields, harvesting and milling wheat, and ultimately "breaking bread" together from the yield. Now, as these volumes go to press, it is nearly time to break bread in celebration of what has been achieved. We look forward to gathering—in person or in spirit—to enjoy and celebrate the harvest of ideas represented here. To everyone who has journeyed with us in bringing this Handbook series to fruition, thank you. We hope that the work born of this collective effort will, in turn, nourish further inquiry and innovation in the service of learning for generations to come.

Toward Assessment in the Service of Learning

Edmund W. Gordon

Pedagogical sciences and practice have long utilized educational assessment and measurement too narrowly. While we have leveraged the capacity of these technologies and approaches to monitor progress, take stock, measure readiness, and hold accountable, we have neglected their capacity to facilitate the cultivation of ability; to transform interests and engagement into developed ability. Assessment can be used to appraise affective, behavioral, and cognitive competence. From its use in educational games and immersive experiences, we are discovering that it can be used to enhance learning. Assessment, as a pedagogical approach, can be used to take stock of or to catalyze the development of Intellective Competence. Educational assessment as an essential component of pedagogy, in the service of learning, can inform and improve human learning and development. This Handbook, in three volumes, points us in that direction.

More than sixty years ago, I had the privilege of working alongside a remarkable educator, Else Haeussermann, whose insights into the learning potential of children with neurological impairments forever altered my understanding of educational assessment. At a time when many viewed such children as unreachable or incapable, Haeussermann insisted that their performances must be interpreted not merely to sort or classify, but to understand—and that understanding must inform instruction. Rather than measuring fixed abilities, she sought to uncover the conditions under which each child might succeed. Her lesson plans were not dictated by standardized norms, but by rich clinical observations of how learners engaged with tasks, responded to guidance, and revealed their ways of thinking. Though her methods defied the conventions of test standardization and were deemed too labor-intensive by prevailing authorities, they represented a

foundational model of what I now describe as *assessment in the service of learning;* assessment not as an endpoint, but as a pedagogical transaction—designed to inform, inspire, and improve the very processes of teaching and learning it seeks to illuminate. The lesson I took from Haeussermann was simple yet profound: that assessment should be used not only to identify what is, but to imagine and cultivate what might become. In every learner's struggle, there is the seed of possibility, and our charge as educators is to create the conditions under which that possibility can take root and flourish.

A Vision for Assessment in Education

In recent years, a profound shift has been gathering momentum in educational thought: the recognition that assessment should **serve** and **inform** teaching and learning processes–not merely measure their outcomes. Nowhere was this vision articulated more forcibly than by the Gordon Commission on the Future of Assessment in Education. Convened over a decade ago under my leadership, the Commission argued that traditional testing–focused on ranking students and certifying "what is"–must give way to new approaches that also illuminate how learning happens and how it can be improved. The Commission's technical report, *To Assess, To Teach, To Learn* (2013), proposed a future in which assessment is not an isolated audit of achievement, but rather a vital, integrated component of teaching and learning processes. It envisioned assessment practices that help *cultivate* students' developing abilities and inform educators' pedagogical choices, thereby contributing to the very intellective development we seek to measure. This call to repurpose assessment—to make assessment a means for educating, not just evaluating—sets the stage for the present Handbook series. Since 2020, I have convened a group of leading scholars to advance the Commission's central proposition with urgency and optimism: that educational assessment, in design and intent, must be reconceived "in the service of teaching and learning."

The need for this reorientation has only grown more pressing. Conventional assessments, from high-stakes tests to admissions exams, have long been designed primarily to determine the achieved status of a learner's knowledge and skills at a given point in time. Such assessments can tell us how much a student knows or whether they meet a benchmark, which may be useful for the purpose of accountability and certification. Yet this traditional paradigm reveals little about how students learn, why they succeed or struggle, and what might help

them grow further. As I have often observed, an assessment system geared only toward outcomes provides a point-in-time picture–a static snapshot of developed ability–but does not illuminate the dynamic processes by which learners become knowledgeable, skilled, and intellectively competent human beings. In effect, we have been evaluating the outputs of education while neglecting the processes of learning that produce those outcomes. The result is an underutilization of assessment's potential: its potential to guide teaching, to inspire students, and to support the cultivation of intellective competence–that is, the capacity and disposition to use knowledge and thinking skills to solve problems and adapt to new challenges. To fulfill the promise of education in a democratic society, we must reimagine assessment as a positive force within teaching-learning processes, one that supports intellectual development, identity formation, equity, and human flourishing, rather than as an external judgment passed upon learning after the fact.

From Measurement to Improvement: Re-Purposing Assessment

Moving toward assessment in the service of learning requires candid reflection on the limitations of our prevailing assessment practices. Decades of research in educational measurement have given us reliable methods to rank, sort, and certify student performance. These methods excel at answering questions like: *What has the student achieved? or How does this performance compare to a norm or standard?* Such information is not without value–it can inform policy decisions, signal where resources are needed, and hold systems accountable for outcomes. However, as we refocus on learners themselves, a different set of questions comes to the fore: How can we improve learning itself? How can assessment and instruction work together to help students learn more deeply and effectively? Traditional tests rarely speak to these questions. A test score might tell us that a learner struggled with a set of math problems, but not *why*–was it a misunderstanding of concept, a careless error, test anxiety, or something about the context of the problems? Nor does the score tell us *what next steps* would help the learner progress. In short, status-focused assessments alone do little to guide improvement. They measure the ends of learning but not the means.

By contrast, the vision of assessment espoused by the Gordon Commission and echoed in my volume *"The Testing and Learning Revolution"* (2015) is profoundly educative in its purpose. In this view, assessment is not a mere endpoint; it is part of an ongoing process of feedback and growth. When assessment is woven

into learning, it can provide timely insights to teachers and learners, diagnose misunderstandings, and suggest fruitful paths for further inquiry. It becomes a continuous conversation about learning, rather than a one-time verdict. This shift entails treating assessment, teaching, and learning as inseparable and interactive components of education—a dynamic system of influence and feedback. I describe assessment, teaching, and learning as a kind of troika or three-legged stool: each element supports and strengthens the others, and none should function independently of the whole. A test or quiz is not an isolated exercise; it is a transaction between the student, the educator, and the content, one that can spark reflection, adjustment, and new understanding. In this transactional view, the student is not a passive object of measurement but an active agent in the assessment process. How a learner interprets a question, attempts a task, uses feedback, or perseveres through difficulty—all of these are integral to the learning experience. Assessment tasks thus have a dual character: they both measure learning and simultaneously influence it.

Embracing this dual character opens up exciting possibilities for re-purposing assessment. Consider, for example, the power of a well-crafted problem-solving task. When a student grapples with a complex problem, the experience can trigger new reasoning strategies, reveal gaps in understanding, and ultimately lead to cognitive growth—if the student receives appropriate guidance and feedback. The late cognitive psychologist Reuven Feuerstein demonstrated decades ago that targeted "instrumental enrichment" tasks could significantly improve learners' thinking abilities; importantly, these tasks functioned as assessments and interventions at once. In the same spirit, assessments can be designed as learning opportunities: rich problems, projects, or simulations that *both* challenge students to apply their knowledge *and* teach them something in the process. A challenging science investigation, for instance, might double as an assessment of inquiry skills and a chance for students to refine their experimental reasoning. When students receive scaffolded support (hints, feedback, opportunities to try again), the assessment itself contributes to their development. In this way, assessment becomes a catalyst for learning. It shifts from a static checkpoint to a dynamic, educative experience. Each assessment interaction is an occasion for growth, not just an audit of prior learning.

Re-purposing assessment also calls for expanding the evidence we consider and collect about learning. If our aim is to understand learners' thinking and guide their progress, we must look beyond right-or-wrong answers. We need to examine *process:* How did the student arrive at this answer? What misconceptions were revealed in their intermediate steps? How did they respond to hints or setbacks? Such evidence may be gleaned through clinical interviews, think-aloud protocols, interactive tasks, or educational games that log students' actions. Today's technology makes it increasingly feasible to capture these rich process data. For example, a computer-based math puzzle can record each attempt a student makes, how long they spend, which errors they make, and whether they improve after feedback–yielding a detailed picture of learning in action. An assessment truly "in the service of learning" will tap into this kind of information, using it to formulate next steps for instruction and to provide learners with nuanced feedback on their strategies and progress. In short, we must broaden our view of what counts as valuable assessment data, integrating qualitative insights with quantitative scores to understand and support each learner's journey fully.

Assessment, Teaching, and Learning as Dynamic Transactions

Central to my proposed paradigm is the understanding that assessment is fundamentally relational and contextual. Learning does not unfold in a vacuum, and neither should assessment. Every assessment occurs in a context–a classroom, a culture, a relationship–and these contexts influence how students perform and how they interpret the meaning of the assessment itself. I speak of the "dialectical" relationship among assessment, teaching, and learning. By this is meant that these processes continuously interact and shape one another like an ongoing dialogue. A teacher's instructional move can be seen as a kind of assessment (gauging student reaction), just as a student's attempt on an assessment task is an act of learning and an opportunity for teaching. When we recognize this, assessment ceases to be a one-way transmission (tester questions, student answers) and becomes a two-way exchange–a *transaction.* In this transaction, students are active participants, bringing their own thoughts, feelings, and identities into the interaction. They are not simply responding to neutral prompts; they are also interpreting *what the assessment asks of them* and *why it matters.* In essence, assessment is a conversation about learning, one that should engage students as whole persons.

This perspective urges us to design assessments that are embedded in meaningful activity and closely tied to curriculum and instruction. Instead of pulling students out of learning to test them, the assessment becomes an organic part of the learning activity. For instance, a classroom debate can serve as an assessment of argumentation skills while also providing students with cycles of preparation and feedback regarding how to formulate and defend ideas. A collaborative applied research project can function as an assessment of problem-solving and teamwork, at the same time building those very skills through practice. In such cases, assessment and instruction intermingle; feedback is immediate and natural (peers responding to an argument, a teacher coaching during the project), and students often find the experience more engaging and relevant. The transactional view also highlights the role of relationships and identity in assessment. How a learner perceives the *purpose* of an assessment and their *relationship* to the person or system administering it will affect their engagement. Do they see the test as a threat or as an opportunity? Do they trust that it is fair and meant to help them? These factors can influence performance as much as content knowledge. Therefore, assessment in the service of learning must be implemented in a supportive, trustful environment. It should feel *to the student* like an extension of teaching–another way the teacher (or system) is helping them learn–rather than a judgment from on high. This more humane and dialogic approach aligns with my lifelong emphasis on humanistic pedagogy: education that honors the whole learner, respects their background and identity, and seeks to empower rather than stigmatize.

Embracing Human Variance and Equity

A commitment to humanistic, learner-centered assessment inevitably leads us to confront the reality of human variance. Learners differ widely in their developmental pathways, cultural and linguistic backgrounds, interests, and approaches to learning. I have often described human variance not as a complication to be managed, but as a core consideration and asset in education. Traditional standardized assessments, in their quest for uniform measures, have often treated variance as "noise" to be controlled or minimized. In contrast, assessment in the service of learning treats variation as richness to be understood and leveraged. Every learner brings a unique profile of strengths and challenges; a truly educative assessment approach seeks to personalize feedback and support to those individual needs. This is not only a matter of effectiveness but of equity

and justice. When assessment is used purely as a high-stakes gatekeeper, it has often exacerbated social inequalities–for example, by privileging those who are test-savvy or whose cultural background aligns with the test assumptions, while penalizing others with equal potential who happen to learn or express their knowledge in different ways. By re-purposing assessments to guide learning, we can instead strive to lift up every learner. Each student, whether gifted or struggling, whether English is their first or third language, whether learning in a suburban school or a remote village, deserves assessments that *help them grow.*

To achieve this, assessments must become more adaptive and culturally sustaining. They should be able to accommodate different ways of demonstrating learning and provide entry points for learners of varying skill levels (the idea of "low floor, high ceiling" tasks). They should also be sensitive to the cultural contexts students bring: the languages they speak, the values and prior knowledge they hold, the identities they are forming. An assessment that allows a bilingual student to draw on both languages, for instance, may better capture–and cultivate–that student's full communicative ability. Similarly, assessments can be designed to honor diverse knowledge systems and ways of reasoning, rather than only a narrow canon. When students see their own experiences and communities reflected in what is being assessed, they are more likely to find meaning and motivation in the task. Moreover, such inclusive assessments can play a role in identity formation: they send a message to students about what is valued in education and whether *they* belong. If assessments primarily signal to some students that they are "failures" or "deficient," those students may internalize negative academic identities, which can undermine their confidence and engagement. But if assessments are reimagined to recognize growth, effort, and multiple and varied abilities, students can begin to see themselves as capable, evolving learners. In this way, a re-purposed assessment system supports not only cognitive development but also the formation of a positive learner identity for every student. Ultimately, embracing human variance is crucial to realizing the broader aim of human flourishing. Education is about nurturing the potential of each human being; assessment should be an instrument for that nurture, helping all learners discover and develop their capabilities to the fullest.

Toward a Pedagogical Renaissance: Analytics and Intellective Competence

Realizing the vision of assessment in the service of learning will require innovation and a renewed research agenda–what we might call a pedagogical renaissance in assessment. One promising path I have begun to explore is the development of "pedagogical analyses" as a robust practice in education. Pedagogical analysis refers to the systematic study of how teaching, learning, and assessment interact–using all available data to understand what works for whom and why. With modern technology, we have more data than ever before about learners' interactions (click streams, response times, error patterns, etc.), and powerful analytical tools, including machine learning, to detect patterns in this data. The goal of pedagogical analysis is not mere number-crunching for its own sake, but to generate actionable insights into the learning process. For example, an analysis might reveal that a particular sequence of hints in an online tutoring system is especially effective for learners who initially struggle, or that students with specific background knowledge benefit from a different task format. These insights allow educators and assessment designers to refine their approaches, tailoring them to a wide range of learners–in essence, personalizing assessment and instruction on a large scale. Importantly, this data-driven approach must be guided by sound theory and a humanistic compass: we seek not to reduce learners to data points, but to augment our understanding of their intellective competence and how it grows.

The concept of intellective competence is central here. Intellective competence, a term I coined, denotes the ability and disposition to use one's knowledge, strategies, and values to solve problems and to continue learning. It is a holistic notion of what it means to be an educated, capable person–going beyond the memorization of facts or routine skills. Our assessment systems should ultimately aim to foster *and* capture these broad competencies: critical thinking, adaptability, creativity, and the capacity to learn how to learn. Doing so means designing assessments that pose authentic, complex challenges to students and then analyzing not only whether students got answers correct, but *how* they approached the challenge. Did they show ingenuity in finding a solution? Did they learn from initial failures and try alternative strategies? Such qualities are the hallmarks of intellective growth. By gathering evidence of these behaviors, we align assessment with the real goals of education in the 21st century. Moreover, assessing for intellective competence has the positive side effect of encouraging teaching toward deeper learning, rather than teaching to a narrow test. When assessments value reasoning, exploration, and

resilience, teachers are more likely to cultivate those capacities in their students. In this way, re-purposed assessments can help bring about a richer educational experience for learners–one that genuinely prepares them for lifelong learning and flourishing in a complex world.

Of course, moving from our current assessment paradigm to this envisioned future is a substantial endeavor. It raises important questions for policy, practice, and research. Policymakers will need to broaden accountability systems to value growth and process, not just point-in-time proficiency. Educators will need professional support to use formative assessment strategies effectively and to interpret the richer data that new assessments provide. Researchers must continue to investigate the best ways to design and implement assessments that embed learning, as well as develop valid ways to infer student understanding from interactive tasks and big data patterns. These challenges, while significant, are surmountable. Indeed, around the world we already see glimpses of the possible: innovative formative assessment programs that transform classrooms into collaborative learning labs; game-based assessments that engage children and teach new skills; participatory assessment approaches that involve students in self- and peer-evaluation, building their metacognitive awareness. Such examples are heartening "existence proofs" that assessment can be reimagined to the benefit of everyone. The task now is to build on these successes, knitting them into a coherent approach that can be implemented broadly and equitably.

The Journey Ahead–and the Contributions of this Handbook Series

This Handbook for Assessment in the Service of Learning series stands as a timely and essential contribution to this educational renaissance. Across its volumes, a breadth of perspectives is presented, all converging on the central theme of transforming assessment to better support teaching and learning. The chapters compiled here bring together renowned scholars and practitioners from a wide range of fields, including cognitive science, psychometrics, artificial intelligence, learning sciences, curriculum and learning design, educational technology, sociology of education, and more. Such range is intentional and necessary. Rethinking assessment is a complex endeavor that benefits from multiple lenses: theoretical, empirical, technological, and practical. Some contributions explore foundational theoretical frameworks, helping us reconceptualize what assessment is and *ought to be* in light of contemporary knowledge about how people learn.

Others delve into the design of innovative assessments, offering design principles and prototypes for assessments that measure complex competencies or integrate seamlessly with instruction. We also encounter rich case studies and practical exemplars–from early childhood settings to digital learning environments–that demonstrate how assessment for learning can be implemented on the ground. These range from classrooms where teachers have successfully used formative assessment to empower students, to large-scale programs that blend assessment with curriculum, to cutting-edge uses of data analytics and AI solutions that personalize learning experiences. The wide-ranging nature of these examples underscores a crucial point: *assessment in the service of learning is applicable in a significant range of educational contexts*. Whether in formal preK–12 schooling, higher education, workplace training, informal learning, or through media and games, the principles remain relevant–aligning assessment with growth, understanding, and human development.

While the chapters in this series each offer unique insights, they are united by a spirit of inquiry, urgency, and hope that echoes the ethos of the Gordon Commission. There is inquiry—a deep questioning of assumptions that have long been taken for granted, such as the separation of testing from teaching, or the notion that ability is a fixed trait to be measured. There is urgency—a recognition that as we move further into the 21st century, with its rapid social and technological changes, the costs of clinging to outdated assessment regimes are too great. We risk stifling creativity, perpetuating inequity, and mis-preparing learners for a world that demands adaptability and continuous learning. But above all, there is hope—a belief that through thoughtful innovation and collaboration, we *can* redesign assessment to be a positive force in education. The work is already underway, and this Handbook is part of it. The range of perspectives in these volumes is a source of strength, encompassing critical analyses, bold experiments, and a blend of longstanding wisdom and fresh ideas, each contributing a piece to the larger puzzle of how to make assessment truly *for* learning.

In closing, let us return to the animating vision that I have championed throughout my career and which inspires this series. It is a vision of education where every learner is seen, supported, and challenged; where assessment is not a grim rite of ranking, but a continuous source of insight and improvement; where teaching, learning, and assessment form a holistic enterprise devoted to nurturing the growth of human potential. Realizing this vision will require perseverance and

creativity. It will mean overcoming institutional inertia and reimagining roles–for test-makers, teachers, students, and policymakers alike. Yet the potential payoff is immense. By making assessment a partner in learning, we stand to enrich the educational experience for all students, help teachers teach more effectively, and advance the cause of equity and excellence by ensuring that every learner receives the feedback and opportunities they need to thrive. This is assessment in the service of learning: assessment that not only reflects where learners are, but actively helps them get to where they need to go next. With the insights and evidence gathered in this Handbook series, we take important steps on that journey. The message is clear and hopeful–it is time to move beyond the extant paradigm and embrace a future in which to assess is, intrinsically, to teach and to learn.

References

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* (Technical report). Educational Testing Service. http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Gordon, E. W., & Rajagopalan, K. (2016). *The testing and learning revolution: The future of assessment in education.* Palgrave Macmillan. https://doi.org/10.1057/9781137519962

Handbook for Assessment in the Service of Learning Series Preface

Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

Objective

How might educational assessment become a catalyst for learning and human development? This question lies at the heart of the *Handbook for Assessment in the Service of Learning* series, Volumes I, II, and III. This series provides a research-based introduction to the theory, design, and practice of assessment in the service of teaching and learning (Gordon, 2020; 2025). The Handbook echoes the call of the *Gordon Commission on the Future of Assessment in Education* to repurpose assessment from merely certifying "what is" to illuminating how learning happens and how it can be improved (Gordon Commission, 2013; Gordon, 2025). The three volumes presented here respond to that call.

Description

The three volumes in this series offer a contemporary view of a range of theoretical perspectives, scholarship, and research and development on innovations with the potential to enable assessment to enhance learning. Across the volumes, contributors explore the central theme of transforming assessment design and development to better support teaching and learning. The three volumes draw on the sciences of learning, measurement, pedagogy, improvement, and more–to inform this charge. We asked authors to anchor chapters in one or more of the design principles for assessment in the service of learning (Baker, Everson, Tucker, & Gordon, 2025). The chapters probe longstanding assumptions, and they explore how to weave a focus on learning into the fabric of educational assessments. The interested reader will find working examples that illustrate what these emerging approaches might look like in practical contexts, from classroom assessments

that empower student agency, to larger-scale assessment systems that, by design, integrate with curriculum and instruction, to applications of data analytics and AI-powered learning platforms that personalize assessment and promote learning. Together, these contributions reflect a common inquiry regarding the design, development, and use of assessment not merely to certify what students know and can do, but to illuminate and support how learning happens and can improve, for every learner (Gordon, 2025; Gordon & Rajagopalan, 2016; Shepard, 2019). From the learner's perspective, well-crafted assessments catalyze and cultivate the very understanding and performance they elicit. Accordingly, the goal is to design educational assessments to nurture productive struggle and growth in the learner.

Audience

This Handbook is intended for a broad audience, from test developers, assessment researchers, and learning scientists to educators, policy makers, and designers. It is a resource for anyone interested in using assessment to help learners learn.

Organization

This Handbook for Assessment in the Service of Learning series is organized into three volumes, each focusing on a critical dimension of assessment in the service of learning. The series includes:

- Volume I: Foundations for Assessment in the Service of Learning
- Volume II: Reconceptualizing Assessment to Improve Learning
- Volume III: Examples of Assessment in the Service of Learning

Together, the volumes present a holistic picture of what it means to redesign assessment in the service of learning—from high-level design frameworks down to concrete tools and practices, and from classroom-level interventions to system-wide exemplars.

Rationale

Too often, assessments have been treated as end-of-learning verdicts—snapshots of what students have achieved—rather than as integral parts of the learning process (Pellegrino, 2014). Meanwhile, important domains of student ability (complex skills like critical thinking and collaboration) have been poorly captured by conventional tests that focus narrowly on easily measured skills (Gordon, 2020).

This Handbook responds to Gordon's charge for assessment innovation. By showcasing successful exemplars, these volumes help define and shape the field that has emerged in the years since the Gordon Commission. Assessment in the service of learning represents a shift in perspective that views assessment, teaching, and learning as inseparable, entangled processes. It envisions a future where every learner is understood, appropriately supported, and sufficiently challenged (Gordon, 1996; Goldman & Lee, 2024). When assessment becomes a partner in the pedagogical aspects of curriculum and instruction, it can enrich and improve teaching and help every learner thrive (Armour-Thomas & Gordon, 2025; Hattie, 2009; Ruiz-Primo & Furtak, 2024). This is the promise of assessment in the service of learning: to not only reflect where learners are, but to actively help them get to where they need to go next. The message of this Handbook is clear: it is time to embrace a future where to assess is to teach and to learn.

References

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of dynamic pedagogy: An integrative model of curriculum, instruction, and assessment for prospective and in-service teachers.* Routledge.

Baker, E. L., Everson, H. T., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 48–92). National Academy of Education.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical Report*. Educational Testing Service.

Gordon, E. W. (2020). *Toward assessment in the service of learning. Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W., & Rajagopalan, K. (2016). *The testing and learning revolution: The future of assessment in education* (pp. 107–146). Palgrave Macmillan.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Pellegrino, J. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116* (11) Article 110313. https://doi.org/10.1177/016146811411601102

Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom activity systems to support ambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The Annals of the American Academy of Political and Social Science, 683* (1), 183–200. https://doi.org/10.1177/0002716219843818.

From Aspiration to Application: Working Examples of Assessment in the Service of Learning

Eva L. Baker, Howard T. Everson, and Eric M. Tucker

Building on the vision articulated in the Series Introduction (Gordon, 2025), this volume answers the call to bridge the chasm between the aspiration for assessment in the service of learning and its practical application. It moves from the 'why' to the tangible 'how' by presenting the 'actionable blueprints' Gómez (2014) called for: concrete examples of assessments that support learning. Drawn from contexts as varied as the College Board's AP® Art and Design portfolios and game-based assessments, these examples are aligned with the core design principles outlined in Baker, Everson, Tucker, and Gordon (2025).

A Framework for Analysis: Three Complementary Lenses

To provide context for these examples, we offer a framework of three complementary 'lenses' from the work of Robert J. Mislevy: Assessment as Evidentiary Argument, as a Feedback Loop, and as Social Practice (Mislevy, 2012, 2018; Bell & Mislevy, 2021). This three-part framework provides a lens for analyzing the working examples that follow, complementing the design principles for assessment in the service of learning proposed by Baker et al. (2025).

The sheer variety of the chapters that follow—from youth development programs to widely adopted digital courseware—calls for shared language for analysis. These examples do more than simply illustrate promising directions for assessment; they reveal aspects of the underlying architecture of learning-oriented assessment designs. To fully appreciate the design trade-offs and innovations detailed ahead, Mislevy's framework invites readers to move beyond viewing these chapters as simple narrations and instead engage with them as complex case studies in assessment design, analyzing how each exemplar succeeds, and where it faces challenges, in integrating the interdependent demands of valid evidence (argument), actionable feedback (loop), and authentic context (practice). This analytical approach is essential for synthesizing insights across chapters and understanding how each contributes to a broader vision of assessment in the service of learning.

Assessment as Evidentiary Argument

The first lens reframes assessment not as a simple measurement tool but as a structured, evidence-based argument (Pellegrino, Chudowsky, & Glaser, 2001; Kane, 2013; Mislevy, Steinberg, & Almond, 2003). From this perspective, a student's performance serves as the data used to support an inference or interpretive claim about their knowledge, skills, abilities, or other attributes. This connection is justified by a warrant and its backing (a generalization supported by theory), requiring designers to first articulate their claims and then construct tasks to elicit the necessary evidence to support those claims.

Assessment as a Feedback Loop

The second perspective shifts focus from the quality of evidence to its use, emphasizing that the data's value depends on how well it informs subsequent decisions. This logic, therefore, requires designers to consider who needs the assessment information, when, and in what form (Hattie & Timperley, 2007; Shute, 2008). It also exposes the tension between a teacher's need for immediate

instructional feedback (a focused, shorter loop) and a system leader's annual data needs (a wider, longer loop). Because an assessment optimized for one purpose is suboptimal for the other, this logic compels a move toward coherent systems of assessments, each designed for a specific purpose.

Assessment as Social Practice

The third lens allows for viewing assessment as a social and instructional activity. Drawing from a sociocognitive perspective, it recognizes that assessments are not neutral instruments but powerful cultural practices that signal what is valued and shape classroom interactions (Shepard, 2000; Bennett, 2023; Nasir, Lee, Pea, & McKinney de Royston, 2020; Penuel & Watkins, 2019). This logic pushes for authentic assessments that mirror real-world disciplinary practices, blurring the line between learning and assessing so that the assessment itself becomes a meaningful learning experience (Mislevy, 2012; Bell & Mislevy, 2021). This perspective also brings issues of human variation and equity to the forefront (Gordon, 1995). It aligns with Gordon's (2020) assertion that designing assessments to respect learners' varied backgrounds and cultivate their abilities is a moral and civil rights imperative. This imperative is a through-line in the chapters that follow, which feature assessments designed for a broad range of learners, from young children interacting with educational media to middle years students developing foundational reading skills.

The Integrated Architecture of Learning-Oriented Assessment

These three perspectives are complementary not separate; together they define the architecture of learning-oriented assessment. The exemplars in this volume show that the promise of innovation rests not primarily on emerging technology, but on the thoughtful integration of their forms of reasoning about assessments intended to support learning. An assessment's capacity to improve learning depends on its ability to elicit valid evidence, provide useful and actionable feedback, and situate itself meaningfully in the social context of teaching and learning (Darling-Hammond, Herman, Pellegrino, Abedi, Aber, Baker, Bennett, Gordon, Haertel, Hakuta, Ho, Linn, Pearson, Popham, Resnick, Schoenfeld, Shavelson, Shepard, Shulman, & Steele, 2013; Goldman & Lee, 2024).

Conclusion

This volume's tangible examples, from badges as assessments to standards-aligned tests and assessments, are offered not as fully formed solutions but as invitations to reflect, iterate, and build upon. They provide the field with a set of powerful existence proofs, hopefully inspiring and better equipping test developers, researchers, and educators to construct more coherent, learner-centered assessment systems that genuinely promote learning and achievement for all learners.

References

Baker, E. L., Everson, H. T., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Baker, E. L., & Gordon, E. W. (2014). From the assessment of education to the assessment for education: Policy and futures. *Teachers College Record, 116*, 1–24.

Bell, C., & Mislevy, R. (2021). Practice, feedback, argument, measurement: A frame for understanding diverse perspectives on teaching assessments. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of analyzing teaching quality: Potentials and pitfalls* (pp. 21–52). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-01

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Darling-Hammond, L., Herman, J., Pellegrino, J. W., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, L. A., Shulman, L., & Steele, C. M. (2013). *Criteria for high-quality assessment* [Technical Report]. Stanford Center for Opportunity Policy in Education.

Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 48–92). National Academy of Education.

Gómez, L. M. (2014). The Gordon Commission: An opportunity to reflect. *Teachers College Record, 116*, Article 110301.

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical Report]. ETS. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 360–372.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Mislevy, R. J. (2012). *Four metaphors we need to understand assessment*. (Commissioned paper for The Gordon Commission on the Future of Assessment in Education). Educational Testing Service.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning*. Routledge.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press. https://doi.org/10.17226/10019

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *The Annals of the American Academy of Political and Social Science, 683*(1), 201–221. https://doi.org/10.1177/0002716219843249

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Practical Examples of Assessment in the Service of Learning at PBS KIDS

Jeremy Dane Roberts, Jessica Wise Younger, Kelly Corrado, Cosimo Felline, Silvia Lovato

PBS KIDS, United States

Abstract

This chapter presents several case studies spanning over a decade of work to demonstrate how PBS KIDS integrates assessment in the service of learning to support its mission of providing effective educational experiences at scale. One case study focuses on a video game designed to teach forces and motion, using a dynamic leveling system that adapts to individual player needs. A research study compares this system to a static approach on learning outcomes. Another case study explores how gameplay data is used to assess counting and cardinality skills for players, training neural networks to predict scores on the Test of Early Mathematics Ability. A third case examines the measurement of behavioral changes in gameplay over time across several PBS KIDS games, developing indicators and models to estimate skill development. A fourth case highlights a machine learning competition aimed at understanding the relationship between game/video engagement and performance on interactive assessments in the PBS KIDS Measure Up! app. Lastly, a final case describes using A/B testing to optimize game design variants, balancing engagement and learning to maximize impact. Together, these cases demonstrate the value of assessment in the service of learning at PBS KIDS.

Author Note
The contents of this chapter were developed under a grant from the Department of Education. However, its contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. [PR/Award No. S295A200004, CFDA No. 84.295A]

Practical Examples of Assessment in the Service of Learning at PBS KIDS

PBS KIDS is committed to making a positive impact on the lives of children through curriculum-based entertainment with positive role models and content designed to nurture a child's total well-being. PBS KIDS' goal is to serve all children. In this chapter, we provide practical examples of how applying the *Principles for Assessment in the Service of Learning* looks in a real-world, scaled up setting. Specifically, we highlight how PBS KIDS, the number one educational media brand for kids (PBS, 2024), has used assessment in the service of learning to further our mission. The work described represents more than a decade of R&D and innovation in learning analytics and learning engineering, driven by our desire to measure, understand, and improve the impact of our media. We have carried out this work in collaboration with a wide range of talented children's media producers, educational researchers, thought leaders, and funders, including the Corporation for Public Broadcasting, the Ready To Learn Program at the U.S. Department of Education, the WGBH Educational Foundation (GBH), University of California, Los Angeles CRESST (UCLA CRESST), and others.

PBS KIDS wants to ensure the media we distribute to millions of children across the US every month (Google Analytics, 2024; Nielsen NPOWER, 2024) have the effect we intend—a positive impact on the lives of all children. In this chapter we focus on how we assess that positive impact through the interactive educational games PBS KIDS distributes. PBS KIDS games offer kids the opportunity to engage with content from a wide range of curriculum in a variety of ways including exploration, tinkering, scaffolded practice, and assessment-focused interactives. These experiences allow kids to explore concepts, practice, get feedback, express what they know, struggle, demonstrate misconceptions, demonstrate mastery, and more. PBS KIDS games present child-relatable situations and challenges, incorporate learning goals, and model problem solving approaches around developmentally appropriate knowledge and skills. The knowledge and skills targeted are selected specifically to help children succeed in school, future work, and life. Accordingly, the

design of the games (and any integrated game-based measurement) incorporates progress, outcomes, and processes, in ways intended to help the children benefit beyond the screens in their everyday life. As needed, PBS KIDS collects fine-grained anonymous user interaction data as children play games to assess different types of learners' knowledge, how it evolves over time, the role our games play in that change, and how we can maximize that role for our media. In this way, we engage in assessment in the service of learning. Children's safety is PBS KIDS' top priority and for that reason, PBS KIDS never collects personally identifiable information.

To date, game-based assessment at PBS KIDS has demonstrated the power of gameplay data to predict scores on standardized tests (Chung et al., 2016; Choi, Suh, Chung, & Redman, 2021), detect (mis)conceptions (Roberts et al., 2019; Lovato, Felline, & Roberts, 2023), assess scientific thinking (Feng, 2019), estimate skill levels for a variety of targeted learning goals including math (Chung et al., 2016), science (Redman et al., 2020; Redman et al., 2021), literacy (Choi, Park, Feng, Redman, & Chung, 2021), and socio-emotional learning (Choi, Suh, Chung, & Redman, 2021), and even measure learning over time (Redman, Feng, Parks, Choi, & Chung, 2023). This chapter will lay out PBS KIDS' vision for assessment in the service of learning in the context of the PBS KIDS mission, audiences, and scale. We provide real-world examples including individualization, assessing skills and measuring impact at scale, and optimizing impact that reflect the following *Principles*:

3. Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

2. Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.

7. Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

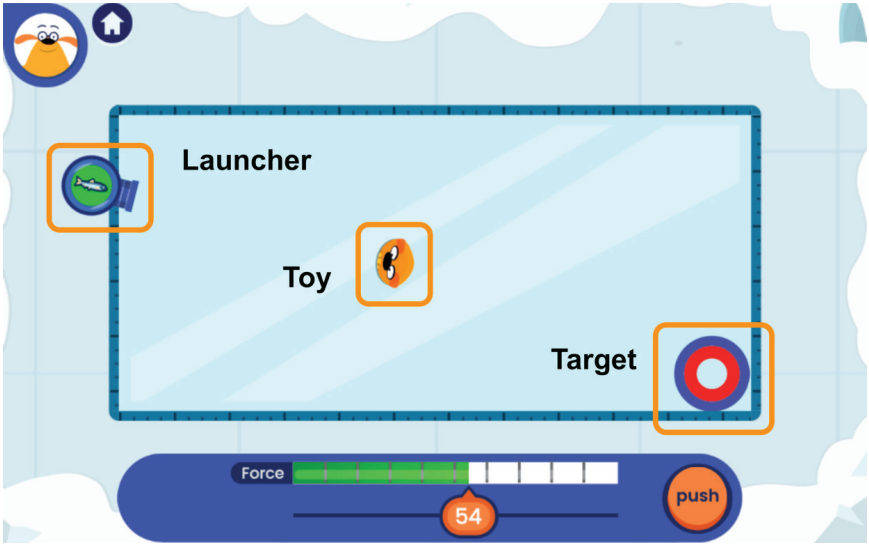
1. Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

Individualization

PBS KIDS recognizes that not all learners have the same needs. We make great efforts to design content and related measurement properties that work well for as many children as possible. This includes a focus on Universal Design for Learning (a research-based educational framework that guides the development of flexible learning environments and learning spaces that can accommodate individual learning differences; Rose, 2000) to guide design decisions such as avoiding requiring background knowledge, experience, or reading ability that is not necessary. To serve a diverse set of learners requires a diverse set of offerings designed to meet learners where they are. To achieve that, we must assess each player. If we can learn about what an individual knows and doesn't know, what they are struggling with or misunderstanding, then we can use that information to make experiences that respond appropriately and adjust to each individual. PBS KIDS believes that game-based assessment can help power individualized learning experiences, in line with *Principle 3: Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.* In the following example, we show how gameplay can be used to estimate a player's skill level and customize their experience by selecting the best next game challenge. Even though this approach did not result in greater learning, it allowed us to understand to what extent a dynamic individualized pathway through a game's levels compares to a static pathway.

As described in Rodriguez, Arena, and Roberts (2018), *Fish Force* is a game that was produced along with videos and activities for the series *The Ruff Ruffman Show* by GBH. *Fish Force* was designed to teach children ages 4–8 concepts of force and motion, like how pushes can have different strengths and cause objects to move in various directions, and how objects can push one another when they touch or collide. Additionally, it was designed to support children in practicing inquiry skills such as making and testing predictions, planning and conducting simple investigations, and engaging in cause-and-effect observations. Players are challenged to rescue a toy plushie stuck on an ice rink by launching a frozen herring at the plushie to knock it onto a target. During the course of the game, players can control the force and/or trajectory of the launcher to attempt to move the plushie to the target area (See Figure 1). Challenge increases between different game levels when additional obstacles are added to the rink—watch out for all of the penguins in the way, ice holes, patches of sand and more! *Fish Force* can be accessed at the PBS KIDS website. (https://pbskids.org/ruff/games/fish-force).

Figure 1.
Example of Fish Force game challenge.



Note. Users can adjust the force meter and the placement of the Launcher to shoot a fish at the Toy to get it to land on the Target while avoiding obstacles. Adapted from Feng, T. (2019). *Using game-based measures to assess children's scientific thinking about force.* [Poster session]. American Educational Research Association Conference, April 5–9, 2019, Toronto, Canada.

In total, 256 game challenges of varying difficulty were created by the *Fish Force* development team, including 128 performance levels (in which the goal is to push the toy to the destination) and 128 prediction levels (of which there are two types: predict the toy's path, or predict where the toy will end up). PBS KIDS games are designed to capture kids' attention, motivate kids to engage deeply, and to be fun so kids invest effort into their play. We theorized we could keep players more deeply engaged by providing them challenges within their zone of proximal development (Vygotsky, 1978). By optimizing engagement, the intent was to promote increased learning outcomes by increasing the amount of instructional material players encountered (Rodriguez, Arena, & Roberts, 2018). That is, rather than provide all learners with the same progression through levels, the game would adapt to support each learner's processes on an individual basis, guiding each player toward the content that would keep them engaged, attentive and motivated, and provide a fun environment to elicit effort to overcome the game's challenges.

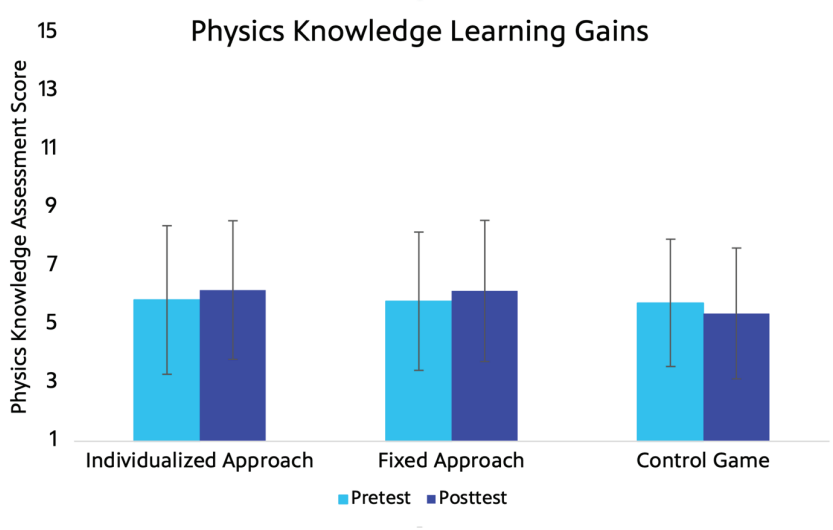
To develop methods for providing an individualized experience, PBS KIDS worked with Kidaptive, a company specialized in individualized learning. Kidaptive applied a Bayesian Item Response Theory (IRT) analysis to rank the difficulty of each game challenge based on an initial sample of players. This model was then incorporated into the game to estimate players' skill levels on the different level types (performance vs prediction) in real time as gameplay proceeded. Similar to computerized-adaptive testing, players' skill estimates were updated after each challenge, and the game used these evolving skill estimates along with the challenge difficulty estimates to select an appropriate next game challenge for the player. Specifically, the probability that the player would correctly solve the next challenge was targeted to be 70%, based on the players' skill level and the challenge difficulty level.

To assess the utility of a personalized approach to gaming, Redman et al. (2019) conducted a study to assess the impact of level progression design on physics knowledge. Students were randomly assigned to play a control game or *Fish Force* with either an individualized level progression (Individualized Approach) or a fixed level progression (Fixed Approach) designed by the game's lead designer and developer. Students were assessed on separate (non-game-embedded) external assessments of children's knowledge of force and motion concepts before and after playing their assigned game. The results (See Figure 2) showed students in both groups that played *Fish Force* made larger learning gains than students

who played a control game. However, the size of the gains was roughly equivalent between the individualized and fixed progression methods.

Figure 2.

Performance on a physics knowledge assessment before and after interacting with the game



Note. Players were assigned to play *Fish Force* with the adaptive level sequence (Individualized Approach), *Fish Force* with the fixed level sequence (Fixed Approach) or a non-physics game (Control Game). The Individualized and Fixed Approach groups showed similar results after controlling for pretest scores. Adapted from Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Schenke, K., Parks, C. B., Michiuye, J. K., Chang, S. M., & Roberts, J. D. (2021). Adaptation evidence from a digital physics game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 55–81). Routledge.

This study demonstrated the feasibility of using assessment to support learners' processes, motivation, and engagement in the context of an educational game. The individualized *Fish Force* game that used game-player data to adapt the game in real time was successful in teaching players physics knowledge. However, implementing the individualized approach did not result in significantly greater knowledge gains compared to the static, fixed approach. This finding suggests personalization may not be required for an assessment to be engaging and motivating. It is possible to create effective media children are motivated to engage with without the costs associated with game-specific development to incorporate real-time skill estimation and adaptive leveling.

As a result of these findings, PBS KIDS is now exploring personalization approaches at a broader and ultimately more scalable level. Real-time adjustments to the levels presented to a player within a single game do not necessarily follow *Principle 7. Feedback for the players to clearly address decisions and next step*s. Therefore, instead of personalization within a single game, we are conceptualizing potential approaches to respond to individual needs when selecting items to engage with from the extensive PBS KIDS media library. By incorporating individualization at the library-level (i.e., a recommendation engine), resources could be focused on a small number of strategically representative games that measure skill level for a variety of learning goals. The player-specific information can then be used to guide the overall learning journey for a player. Such an approach would also better align with *Principle 7* by providing clearer next steps for a player to build on their current skill set via the suggested content. This library-level approach may provide higher quality individualization by not only keeping a player engaged at the appropriate challenge level across the media they engage with but also suggesting related or new content to encourage diversifying the topics learned.

Assessing Skills at Scale

In addition to assessing individuals, we believe assessment of our audience has the potential to answer important questions about young children at the group level. The millions of monthly users PBS KIDS games reach (average of 3.4 million unique monthly users on the PBS KIDS Games app and 6.9 million on pbskids.org; Google Analytics, 2024) represents a sizable sample of children aged 2–8, a population that has historically been expensive and difficult to measure systematically, particularly in naturalistic settings (Nagle, Gagnon, & Kidder-Ashley, 2020). As such,

it has been difficult to assess what young children know (their prior knowledge) to understand their educational needs. For example, what are children's skill levels across various subjects, where are their needs greatest, and what are the implications for investments in new educational content? While the United States tracks such information starting in 4th grade through the National Assessment of Educational Progress, no such program exists for preschool, in part due to the difficulty of assessing children this age at scale. This lack of insight into young children's knowledge represents a gap in understanding of kindergarten readiness and the resources needed to support our youngest learners. PBS KIDS believes by designing game-based assessments that meet *Principle 2. Assessment that focuses on progress, outcomes, and processes that can be transferred to other settings, situations, and conditions,* we can inform PBS KIDS' curriculum focus over time to meet demonstrated needs in particular areas. For example, if a particular skill set sees a dip in performance, PBS KIDS can adjust production to develop more related media or better promote and make more discoverable existing content that responds to the need. Below, we discuss an example that demonstrates a proof of concept for such population-level assessment of children's knowledge via gameplay data.

Curious George Busy Day is a set of 16 games, available in English and Spanish, that were developed by GBH, and that focus on counting and cardinality. The set of 16 games represent learning goals such as number knowledge and counting skill. Three games, *Apple Picking*, *Blast Off*, and *Meatball Launcher*, have game mechanics that require players to make a judgment about numbers and actions and therefore can be used to assess player skill level. Specifically, *Apple Picking* assesses a player's ability to count on by ones from a number other than 1 by requiring players to select the missing number in a sequence. *Blast Off* assesses the ability to count backwards from 10 by asking players to select a series of numbers from largest to smallest. Finally, *Meatball Launcher* assesses the ability to count or put out 1 to 5 objects upon request by asking players to give a requested number of items. These tasks are illustrated in Figure 3 and can be accessed at the PBS KIDS website. (https://pbskids.org/curiousgeorge/busyday).

Figure 3.
Example game challenges from Curious George Busy Day

Note. In Apple Picking (A) players must select the apple with the number that belongs where the question mark is in the line of apples. In Blast Off (B), players must select the numbers from largest to smallest to blast off the rocket. In Meatball Launcher (C), players must put the requested number of meatballs on the plate.

As described in Roberts et al. (2018), researchers at UCLA CRESST first conducted analyses examining whether measures of game progress (rounds completed, time spent, time to correct answer), game performance (number of correct first attempts, number of overall correct attempts, number of overall incorrect attempts), or their combination were related to scores on a standardized assessment, the Test of Early Mathematics Ability, 3rd Edition (TEMA-3; Ginsburg & Baroody, 2003). Generally, performance-based measures were more strongly related to test scores than progress-based measures. Across all three games, the strongest positive predictor of math knowledge was the number of correct first attempts at a solution, while the strongest negative predictor was the number of incorrect solution attempts. However, measures that incorporated both progress and performance yielded the highest correlations with the TEMA-3. Specifically, vector combinations that incorporated success in one dimension, but error in another; number of first correct attempts (success) and time taken to correct first attempt (error) ranged from 0.43 to 0.58 and number of incorrect attempts (error) and highest level reached (success) ranged from 0.48 to 0.76 across all three games (See Table 1). Interestingly, *Meatball Launcher* consistently had strong correlations with TEMA-3 scores across all measures. This game was the only examined game that did not provide feedback as to the accuracy of the answer. This finding suggests children do incorporate in-game feedback into their gameplay and can learn from the test. For PBS KIDS, the implication is that if we wish to assess our audience's skill level, some games should be designed solely for assessment purposes (not a hybrid of instruction and assessment) to provide a more accurate measurement.

Table 1.

Correlations (Spearman) Between Vector-Based Angular Component Measures and TEMA-3 Measures by Game

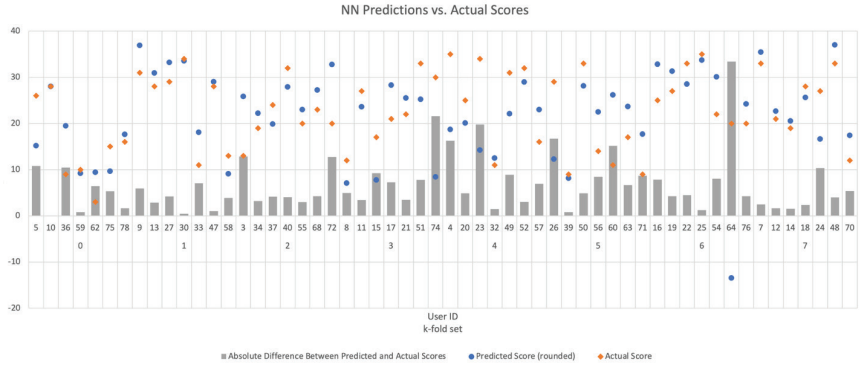
Measure	Total score	Cardinality subscale	Counting subscale
Vector 1: y = No. of correct first attempts, x = Time taken for first attempts (min.)			
Apple Picking	.43**	.37**	.35**
Blast Off	.51***	.40**	.41**
Meatball Launcher	.58***	.52***	.50***
Vector 2: y = No. of correct attempts, x = Mean level time (min.)			
Apple Picking	.26	.20	.12
Blast Off	.42**	.34*	.30*
Meatball Launcher	.70***	.63***	.61***
Vector 3: y = No. of incorrect attempts, x = Mean level time (min.)			
Apple Picking	-.28*	-.26	-.32*
Blast Off	-.28	-.19	-.23
Meatball Launcher	-.38*	-.40*	-.34*
Vector 4: y = No. of correct attempts, x = Highest level reached			
Apple Picking	-.35*	-.34*	-.23
Blast Off	-.48***	-.43**	-.40**
Meatball Launcher	.09	.13	.13
Vector 5: y = No. of incorrect attempts, x = Highest level reached			
Apple Picking	-.48***	-.41**	-.36**
Blast Off	-.57***	-.52***	-.51***
Meatball Launcher	-.76***	-.71***	-.64***

*$p < .05$ (two-tailed). **$p < .01$ (two-tailed). ***$p < .001$ (two-tailed).

Note: From Chung, G. K. W. K., & Parks, C. (2015b). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). University of California, National Center for Research on Evaluation, Standards, and Student Testing.

UCLA CRESST then examined how more game-based information about a player might be used to improve predictions on their standardized test performance (Chung & Parks, 2015b). Using 1702 different indicators derived from seven different games from *Curious George Busy Day*, UCLA CRESST built and trained several neural network models. The models were each trained on one subset of data, then validated on another subset. The best-performing model that leveraged data from many indicators of skill across seven games on average predicted individual's TEMA-3 scores within about 8% of their actual score (See Figure 4).

Figure 4.
Neural Net (NN) TEMA-3 predicted and actual scores



Note: Adapted from Roberts, J. D., Parks, C. B., Chung, G. K. W. K., Redman, E. J. K., Schenke, K., & Felline, C. (2018). Innovations in evidence and analysis: The PBS KIDS Learning Analytics Platform and the research it supports. In *Getting Ready to Learn* (pp. 231–248). Routledge.

The results of this study demonstrated the very real potential to use games as assessments that meet *Principle 2: Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.* Performance on the selected *Curious George Busy Day* games relates to a completely different and meaningful context: performance on the TEMA-3 standardized test. This work shows that such assessment can be done at scale with young children and demonstrates PBS KIDS is capable of performing benchmarking at the population level through our games.

Measuring Impact at Scale

PBS KIDS serves the American public at scale, and desires to measure the impact we make with media at scale. We define impact as a combination of reach, engagement, and learning effectiveness. For a child to learn something from PBS KIDS media, we must reach them, they must choose to engage, and the media must be effective at promoting learning. PBS KIDS believes that in-game assessments can help us measure the learning component of our impact by following *Principle 7: Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.*

Measuring reach (number of users exposed to our content) and engagement (how long a user engages, amount of content engaged with, etc.) are relatively straightforward. Measuring learning effectiveness is much more difficult. Unlike reach and engagement, which can largely be measured by counting users and their interactions, effectiveness implies measuring a change in users' performance over a period of time. Historically, large scale randomized control trial (RCT) studies have provided such information on PBS KIDS media. However, these efforts have limitations, particularly when considering employing them at scale. While still considered the 'gold standard' for determining the instructional potential of specific pieces of media, these RCTs are slow and expensive, and do not always reflect how the content is used "in the wild" (Redman et al., 2021). These limitations result in RCTs being conducted on only a small subset of content and leave the effectiveness of the media when used under typical, unguided conditions unclear. Specifically, RCT participants are often directed to use the material in a prescribed, consistent way over a period of weeks, and the material is often in isolation from any other PBS KIDS offerings. However, in non-research settings, users interact with the same game content from within a much larger suite of media offerings (as of 2024, the PBS KIDS Games app offers almost 300 games), and engagement patterns can differ substantially. For one studied set of games, less than 1% of the PBS KIDS Games app population engaged with the games to a similar depth of content coverage as the recruited study population (Choi, Suh, Chung, & Redman, 2021), signaling a potential lack of effectiveness for our population. However, this comparison between populations did find support for the generalizability of efficacy of our content for our population, as gameplay performance and the skill level estimates from psychometric models were similar between the recruited study sample and those players that engaged at a similar level in natural settings.

In the hopes of measuring learning effectiveness faster, more cost effectively and with a more naturalistic sample, PBS KIDS and UCLA CRESST set out to develop a way to use gameplay data from the PBS KIDS audience to directly measure changes in behaviors that are consistent with a player learning over time. Further, to maintain children's privacy, this work had to be conducted with anonymous gameplay data and not incorporate any demographic information. Such an endeavor would extend previous work aimed at assessing an individual's skill at a single point in time (e.g., Roberts et al., 2018; Roberts et al., 2019) to follow how that individual's skill differed across multiple timepoints. While a logical and relatively straightforward extension of previous work, this project presented new challenges. Specifically, we sought to understand whether the changes in behavior could be reasonably attributed to a player's interaction with the PBS KIDS game without knowledge of activities done outside of their interactions with PBS KIDS games. However, if successful, the work could be used to develop an indicator of learning effectiveness that is consistently monitored and reported on, similar to the metrics used for reach and engagement.

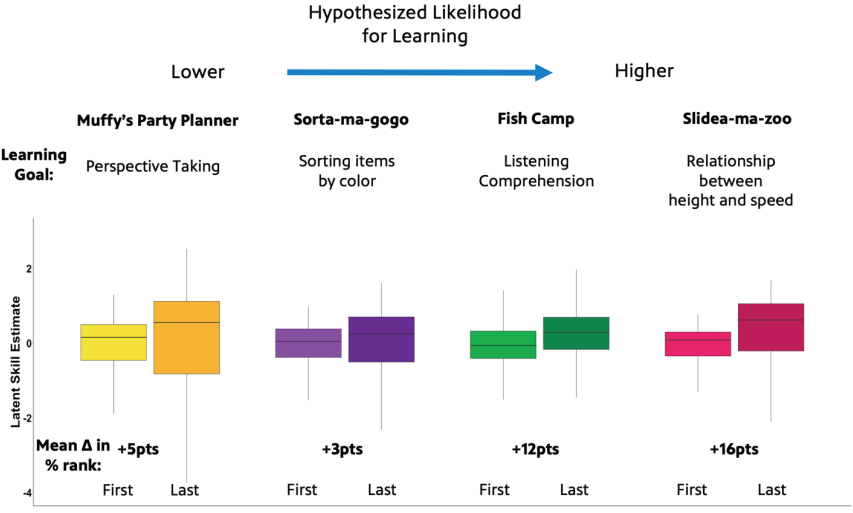
As part of the initial effort at measuring learning over time, Redman et al. (2023) first selected a subset of PBS KIDS games from which skill level at a given construct could be reasonably estimated using gameplay data alone. These games were then evaluated for the potential to promote learning based on features of the games, such as whether user feedback was provided or constructed learning (Nanjappa & Grant, 2003) was encouraged. This evaluation was called a "qualitative ratings validation approach". The four games included in the final analysis represented a range of potential for learning. Specifically, based on the availability and quality of feedback mechanics (incorrect answer elaboration, graduated feedback) and constructive learning processes (prediction, reflection, and debugging/correction), *Slidea-ma-zoo* from the series *The Cat in the Hat Knows a Lot About That!* and *Fish Camp* from *Molly of Denali* were designated as having a high potential for learning. The game *Sorta-ma-gogo* from *The Cat in the Hat Knows a Lot About That!* did not have as much elaborative feedback or encourage player reflection and so was rated as having less potential for learning. Finally, *Muffy's Party Planner* from the series *Arthur* was specifically designed to measure and not teach. Therefore, there was no feedback or constructive processes involved in the game, and it was rated as low potential for learning.

The inclusion of *Muffy's Party Planner* designed for measurement only was key for our validation process. Namely, by examining games with both low- and high-likelihood of learning, we could assess how likely skill gains were due to engagement with the PBS KIDS games. Indeed, young children should be improving their skill sets over time through a variety of opportunities in their daily life, so learning gains not specific to interactions with the game were expected. This qualitative ratings validation approach was determined to be faster and more cost effective than implementing a small, recruited study looking at correlations between external measures and gameplay-based estimate of skill. For PBS KIDS, this work represented a novel symmetric approach to simultaneously validate the utility of game-based performance measures as indicators of skill on a construct, the models used to estimate player skill level, and the qualitative rating system for a game's likelihood of learning. It also provided key data on how much confidence we should have in these approaches.

Next, for each game, UCLA CRESST used an IRT model to estimate the difficulty and discrimination parameters of each challenge or 'item' in a game. A player's skill level on a given construct targeted by the game was then estimated at two time points at least one day apart based on the player's responses to game challenges and the item parameters. Skill change score was determined by subtracting the initial estimate from the second estimate. We found that, as expected, changes in skill were detected across all games. Importantly, though, the size of the gains was generally larger for games with higher potential for learning and lower for games with lower likelihood for learning (See Figure 5). Players of both games rated as having high potential for learning showed larger gains in skill estimate over two time points compared to the changes seen in skill estimates of players of *Muffy's Party Planner*, the measurement game. This initial effort took a conservative approach to provide preliminary proof of concept to measure the efficacy of a game using only anonymous gameplay data. Specifically, strict inclusion criteria, including requiring participants to interact with specific game challenges more than once, and at least one day apart, resulted in only about 10% (N=237,293) of the full data set (N=2,174,787) being analyzable in the model of skill change. Further, while the users included in the analysis showed significantly higher engagement with the games compared to those not analyzed, similar to the comparison between recruited study participants and PBS KIDS Games app users at large, initial performance between the included and excluded players was similar. This comparison indicated that the included players likely did not have greater initial skill compared to the excluded players and suggested the non-studied players would have similar potential for learning gains if they interacted with the game more.

Figure 5.

Latent skill estimates from first and last encounter with a game



Note. Learning gains were roughly consistent with the hypothesized likelihood for learning developed from feature analysis. From Younger, J. W., Roberts, J. D., Felline, C., Corrado, K., & Lovato, S. *The role of learning analytics at PBS KIDS*. [Poster session]. Biennial Meeting of the International Mind Brain and Education Society, July 10–12, 2024, Leuven, BE.

Future work is planned to create models of skill that better reflect the needs of our PBS KIDS audience, namely, constructing valid models for detecting skill change within a shorter period of time that better align with the natural game engagement pattern of our users. Further, we plan to include additional input to the model such as amount of content covered within a session, time between sessions, and more to further refine our models to make better inferences about the effectiveness of PBS KIDS games (Chung, Redman, & Choi, 2023). In this way, as outlined in *Principle 6*, we expect to ensure that our assessment purposes fit our audience, improve the credibility of our assessments, and draw appropriate inferences from them, ultimately helping us succeed in meeting the PBS KIDS mission.

Optimizing Impact at Scale

Beyond measuring impact, another key use case for assessments for PBS KIDS is to continuously improve our approaches toward impact over time. To achieve this goal requires following *Principle 1: Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.* More specifically, by understanding the specific engagement patterns of individuals as they play games and measuring learning gains as they play, we can determine the relationship between what players do and what they learn. A clear picture of this relationship will help us develop models of impact and improve them over time. However, not all our efforts to link player behavior and learning help move our models of impact forward. In the following examples, we present cases that demonstrate how the assessment transparency noted in *Principle 1* can directly impact our ability to serve learning.

In 2019, PBS KIDS was part of a competition focused on Artificial Intelligence (AI), and its application to various disciplines and hard problems (Felline et al., 2019). Competitors from across the globe tackled a specific challenge and competed on well-defined scoring criteria. The challenge was to use anonymous interaction data from users engaging with a variety of video and games to predict performance on embedded interactive assessments within the PBS KIDS *Measure Up!* app, which was developed to teach preschool and early elementary school-aged children measurement concepts such as height and length, weight, and capacity. The hope was the competitors would help extend previous efforts showing the app was successful in improving children's knowledge of pan balances (Schenke et al., 2020) to understand why children were likely to benefit.

PBS KIDS viewed this data challenge as an opportunity to understand how AI and machine learning approaches could help discover relationships between engagement with various specific features of the media and outcomes on the assessments. Sophisticated models were able to predict players' game-based assessment performance based on their game interaction data reasonably well. Models were scored on a scale of -1 to 1 using methods of measuring inter-rater reliability between the model predicted scores and the actual scores (quadratic weighted kappa; McHugh, 2012). The winning model achieved a score of 0.568, with 0.6 considered a very good score. However, the winning teams employed models that could not be fully explained to PBS KIDS. As such, there was no way to gain insights into the media design choices in the studied games to enable these

predictions to be applied in other scenarios or even iterated on in a theory-driven way. PBS KIDS considers the limitations of such models to be serious enough that we have shifted our focus almost exclusively to explainable models. Without assessment transparency, PBS KIDS cannot improve our models for impact.

PBS KIDS is now taking a different approach to obtain the assessment transparency needed to power the continuous improvement of the design of educational media. We are now conducting randomized control trials directly within a PBS KIDS flagship distribution product: the PBS KIDS Games app. Our approach is to have each experiment-capable game incorporate several variable experiences that can each be independently manipulated. In this way, many different aspects of game design and their potential interactive effects can be examined within the same context of a given game design. Maximizing the experimental space of a given game also allows us to conduct fast analytics-based randomized control trials at the scale of the PBS KIDS Games app audience. At the time of writing, we have completed experiments on two different games that each have the target goal of teaching players about the design process, though designed with different age groups in mind. As explained in Younger et al. (2024), both experiments examined how the level of specificity of game instructions impacted player behavior (Mayer, 2023). The first experiment with 1,054,651 enrolled users additionally examined the effect of prompt construction, comparing a question vs statement format (King, 1991). The second experiment with 567,267 users additionally examined the impact of motivational elements in the game. Through these experiments, we were able to identify which elements of a game are likely to be most impactful to players' experience. In the first experiment, the instruction specificity variable was manipulated within instructions that were verbal in nature (either read by or spoken to the player by in-game). The variable was implemented in two different phases of the game with the intent to determine whether specificity would be more impactful at different phases of learning or whether there may be additive effects (e.g., two specific instructions might be more impactful than one). Yet, there were no meaningful differences across our different experimental conditions. Indeed, as many as 30% of users chose to skip the instructional prompt with the experimental manipulation, though these users did not perform differently from those that did not skip the instruction. We hypothesized multiple explanations for this finding. First, the timing of the specific instructions relative to expected user actions may not have been appropriate to impact user behavior. Second, the presence of additional supportive visual elements present in the game at the time

of instruction were more salient to players than the verbal instructions presented. The transparency of our assessment methods allowed us to iterate on these ideas in future experiments. In the second experiment, we adjusted the manipulation such that it took place in earlier, initial instructions to the user that were visual in nature rather than verbal. In this experiment, the different variable conditions did produce meaningful differences in user behaviors in the game. Those players that received more specific instruction were more likely to make use of features in the game designed to aid performance and required fewer attempts to complete the challenges presented in the game compared with users who received less specific instruction.

While analysis can identify which variants might be most effective for learning, multiple lenses are required to determine the overall impact of a variant. As mentioned earlier, the informal media landscape is filled with many activities for kids to choose to engage with. It is therefore not enough for an experience to be highly educational alone. If kids choose to engage in something else and engagement with our media goes to zero, then impact also goes to zero. Therefore, in addition to comparing how variables influence learning, we consider whether they affect engagement. For example, although we may have chosen to make the instructional prompts non-skippable, through prior work, we know that engagement with a game tends to drop if instructions are required before users can interact with the game. Therefore, while a variable might influence how many attempts it takes a player to solve a particular challenge, we must also ensure players are engaging with the same number of challenges across all experimental conditions. What is the proper balance between engagement and effectiveness? In the experiments run to date, there were no differences in engagement across experimental groups. However, as we expand our experiment program to different types of variables, it is our hope to establish a quantitative understanding of the balance between engagement and effectiveness. Ultimately, this foundation will support team debate, definition, and alignment toward a quantitative definition of impact itself and how impact is aggregated across millions of users and relevant subgroups. As we establish a baseline understanding of what is true today, we will use this understanding to help us improve going forward. Developing the capability to discover the optimal design principles of educational games will provide the feedback that game producers, designers, and developers need to help make decisions about how to proceed with game development iterations, and with future game design efforts.

Progress and Implications

PBS KIDS has been fortunate to develop and execute a variety of projects that all focus on using assessment in the service of learning. This program of work has required over a decade of systematic work across children's media producers, educational researchers, thought leaders, and funders to innovate the tools, technology, and processes needed to measure, understand, and improve PBS KIDS games. First and foremost, the game-based assessment work would not be possible without data collection infrastructure. Over the years, PBS KIDS developed a bespoke system for data collection to meet the many needs for our research program. We capture very detailed anonymous interaction data (which includes no personally identifiable information) from PBS KIDS games. Data collected by our system includes events capturing time series data around user action, system reactions, instruction, feedback, hints, voice over captions, and snapshots of the evolving state of game challenges such as puzzles and problem-solving tasks. As such, much more data are generated from our system compared to more typical business use cases aimed at understanding user activity. Therefore, as we collected more data from more sources across games and distribution channels, we developed tools for great control over when and where data are collected. This high degree of control has the dual benefit of supporting both privacy and sustainability goals. Other important steps to scaling data collection include standardizing log data across games to allow for greater consistency and efficiency of analysis and the game development process itself. For example, PBS KIDS has certain requirements for games distributed on its platform. By fitting our data collection platform into this ecosystem, we could more easily ensure all games commissioned by PBS KIDS have the potential to use our system if desired.

Our approach to data collection leads to interesting limitations in the data collected such as the absence of information about a player's background, demographics, specific setting, and a lack of knowledge of whether a single device is being shared amongst multiple individuals during co-play. Despite these limitations, as the examples above show, the data power research that is safe and valuable. Further, to supplement the large-scale anonymous data collected, PBS KIDS also commissions recruited studies that can collect additional demographic data through formal research consent processes. A series of tools (e.g., to easily deploy games into research environments, configure data collection, and provide researchers with easy access to study data) were created to facilitate these

studies and enable PBS KIDS games to be researched in more controlled settings and ensure research data is separate and distinct from that collected from the general population.

Another equally important contributor to the success of our research program has been the cultivation of data awareness and use of gameplay data, assessment, and the related potential for measuring and optimizing impact. We have strived to amplify the results of our work both internally to product development and strategy teams and externally to academic and industry groups. Meeting these goals has required research agendas that are developed in a mutually beneficial fashion, contributing to both foundational work around the potential to use game-based assessment for learning as well as more immediate tangible benefits to the wider PBS KIDS community. For example, during launches of new games, highly detailed user interaction data are collected with the intent of understanding how to measure learning from player behavior. These same data can be used to understand important player patterns such as where players may encounter unexpected difficulty with the game, which can be reported back to the game developers who can adjust the game as necessary. Building such symbiotic research programs has emphasized the importance of individualizing our approaches to learning and teaching within our own team, and across the community of production partners. Just as we develop different games to meet the needs of different learners, we have had to evolve our research programs to meet the needs of different consumers of our work. Adapting to meet the needs of our consumers has resulted in developing analytic pipelines that can operate on different time scales. An academic pipeline, for example, might take place on a longer time scale and include detailed statistical analysis presented in a formal report. A game development pipeline, on the other hand, may operate on a much faster scale, taking samples of data and using visualizations to quickly assess whether a feature seems to be working or not. This allows data-informed iteration and improvement to be seamlessly integrated into our development processes, which is considered vital to PBS KIDS and our digital producers. Communicating insights in a way that is familiar and approachable for different audiences has been instrumental to growing our support base, and therefore our research program capabilities. There is much more for us to explore around how best to support the collaborations and processes that power the development of PBS KIDS games, distribution platforms, user experiences, marketing and promotional strategies, distribution strategies and more.

Looking forward, we hope to continue our efforts related to assessment in the service of learning on multiple fronts. First, we want to continue to innovate on how we develop and validate new models for assessment. This effort includes continuing to improve how we determine whether models are suitable for the purposes for which we create them. In particular, we want to ensure the inferences we make and the decisions we take based on them are aligned with our objectives. Next, we want to expand our effort to support learner's processes with individualized instruction in ways that encompass the larger PBS KIDS library of media, including both games and videos. We are currently in early exploration and planning around recommendation engines, and how they can be applied appropriately in the PBS KIDS context and expect to learn a lot over the next few years. Finally, we want to further demonstrate that the skills players exhibit while playing PBS KIDS games (such as the *Curious George Busy Day* games) can transfer to other different and important contexts beyond the TEMA-3, e.g., on performance tasks in the real world.

After over a decade of work and a variety of principles coming together, collectively we have accomplished much. We have developed large-scale, high value, and safe gameplay data collection capabilities to power game-based assessment-powered individualized learning approaches, models to estimate skill levels on learning goals using gameplay, and models for estimating learning over time based on the skill estimates. We have further crafted a method for the systematic, speedy, and efficient discovery (and improvement over time) of design principles for educational children's media that work best at scale. What will the next decade bring?

References

Choi, K., Parks, C. B., Feng, T., Redman, E. J. K. H., & Chung, G. K. W. K. (2021). *Molly of Denali Analytics Validation Study Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Choi, K., Suh, Y. S., Chung, G. K. W. K., & Redman, E. J. K. H. (2021). *Population study final study report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Chung, G. K. W. K., & Parks, C. (2015). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G. K. W. K., Parks, C. B., Redman, E. J. K. H., Choi, K., Kim, J., Madni, A., & Baker, E. L. (2016). *PBS KIDS Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Chung, G. K. W. K., Redman, E. J. K. H., & Choi, K. (2023). *Wombats Analytics Evaluation–Final Plan* (Final deliverable to PBS KIDS). UCLA/CRESST.

Felline, C., Roberts, J. D., Rohner, J., Oder, J., Springer, K., Corrado, K., Demkin, M., & Cukierski, W. (2019). *2019 Data Science Bowl*. Kaggle.

Feng, T. (2019). *Using game-based measures to assess children's scientific thinking about force.* [Poster session]. American Educational Research Association Conference, April 5–9, 2019, Toronto, Canada.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). ProEd.

Google Analytics. (2024a). *Analytics 360: PBS KIDS Games app* 20230701–20240630.

Google Analytics. (2024b). *Analytics 360:* pbskids.org (browser traffic only, excluding WebView browsers) 20230701–20240630.

King, A. (1991). Effects of training in strategic questioning on children's problem-solving performance. *Journal of Educational Psychology, 83*(3), 307–317. https://doi.org/10.1037/0022-0663.83.3.307

Lovato, S., Felline, C., & Roberts, J. (2023). *Measuring distance between solutions in an engineering game for children.* [Poster session] Society for Research in Child Development Conference, March 23–25, 2023, Salt Lake City, UT.

Mayer, R. E. (2023). Improving learning from screens for toddlers and preschoolers. *Journal of Applied Research in Memory and Cognition, 12*(4), 473–475. https://doi.org/10.1037/mac0000133

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Nagle, R. J. (2007). Issues in preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 29–48). Lawrence Erlbaum Associates Publishers.

Nanjappa, A., & Grant, M. M. (2003). Constructing on constructivism: The role of technology. *Electronic Journal for the Integration of Technology in Education, 2*(1), 38–56.

Nielsen NPOWER. (2024). L+7, 9/25/23 - 9/29/24, M-Su 6A-6A Reach (000), PBS stations, 50% unif., 1+ min.

PBS. (2024). *2024 PBS Trust Survey* [Flyer]. https://dc79r36mj3c9w.cloudfront.net/prod/filer_public/value-pbs-bento-live-pbs/Downloadables/935e535c5e_PBS%20Trust%20Survey%20Flyer_2024.pdf

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Parks, C. B., Schenke, K., Michiuye, J. K., Choi, K., Ziyue, R., & Wu, Z. (2020). *Cat in the Hat Builds That analytics validation study* (Final deliverable to PBS KIDS). UCLA/CRESST.

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Schenke, K., Parks, C. B., Michiuye, J. K., Chang, S. M., & Roberts, J. D. (2021). Adaptation evidence from a digital physics game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 55–81). Routledge.

Redman, E. J. K. H., Feng, T., Parks, C. B., Choi, K., & Chung, G. K. W. K. (2023). *Learning-related analytics KPI—KPI Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Redman, E. J. K. H., Parks, C. B., Michiuye, J. K., Suh, Y. S., Chung, G. K. W. K., Kim, J., & Griffin, N. (2021). *Social-emotional learning games validity study* (Exploratory study): Final study report. UCLA/CRESST.

Redman, E. J. K. H., Schenke, K., Chung, G. K. W. K., Parks, C. B., Michiuye, J. K., Feng, T., Chang, S. M., & Cai, L. (2019). *Analytics Validation Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Roberts, J. D., Chung, G. K. W. K., Feng, T., Riveroll, C., Redman, E. J. K. H., Schenke, K., Lund, A., & Rodriguez, J. (2019). *Deriving learning-related measures from game telemetry: Detecting children's alternative conceptions of the pan balance.* [Poster session]. Biennial Meeting of the Society for Research in Child Development, March 21–23, Baltimore, MD.

Roberts, J. D., Parks, C. B., Chung, G. K. W. K., Redman, E. J. K., Schenke, K., & Felline, C. (2018). Innovations in evidence and analysis: The PBS KIDS Learning Analytics Platform and the research it supports. In *Getting Ready to Learn* (pp. 231–248). Routledge.

Rodriguez, J., Arena, D., & Roberts, J. D. (2018). Adaptive and personalized educational games for young children: A case study. In *Getting Ready to Learn* (pp. 212–230). Routledge.

Rose, D. (2000). Universal design for learning. *Journal of Special Education Technology, 15*(4), 47–51. https://doi.org/10.1177/016264340001500108

Schenke, K., Redman, E. J. K., Chung, G. K. W. K., Chang, S. M., Feng, T., Parks, C. B., & Roberts, J. D. (2020). Does "Measure Up!" measure up? Evaluation of an iPad app to teach preschoolers measurement concepts. *Computers & Education, 146*, 103749. https://doi.org/10.1016/j.compedu.2019.103749

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (A. R. Luria, M. Lopez-Morillas, & M. Cole, Trans.). Harvard University Press.

Younger, J. W., Felline, C., Killian, R., Corrado, K., & Roberts, J. D., *Evaluating effectiveness of educational games in natural settings*. [Conference presentation abstract]. 2025 Digital Media and Developing Minds International Scientific Congress, July 13–16, 2025, Washington D.C., USA.

Younger, J. W., Roberts, J. D., Felline, C., Corrado, K., & Lovato, S. *The role of learning analytics at PBS KIDS*. [Poster session]. Biennial Meeting of the International Mind Brain and Education Society, July 10–12, 2024, Leuven, BE.

Credits

Privacy

PBS KIDS is committed to creating a safe and secure environment that family members of all ages can enjoy. Children's privacy and safety are our top priority. As such, PBS KIDS never collects personally identifiable information. Consistent with our privacy policy, we and our service providers (like Google Analytics) intend to only collect and analyze data that is needed to deliver the high quality educational media experiences that users expect and to operate necessary business functions. To view the full PBS KIDS privacy policy, please visit: pbskids.org/privacy/.

Assessment in the Service of Learning: An Example from AP® Art and Design

Rebecca Stone-Danahy, David S. Escoffery, Natalya Tabony, and Trevor Packer

With its focus on providing support materials for teachers and students that allow opportunities for real time feedback, the 2019 redesign of Advanced Placement (AP®) courses solidified the AP Program's commitment to the Assessment in the Service of Learning (AISL) ideals. AP Art and Design offers a model for the ways in which assessments can support the process of learning. The Art and Design course and assessment both drive student motivation, engage students in some way, and promote metacognitive skills. This chapter examines the structure of the AP Art and Design portfolio assessment along with the support offered to teachers and students, demonstrating how this program models the process-focused elements of AISL. Because the portfolio requires students to conduct an inquiry emphasizing process over product, AP Art and Design provides inherent motivation for students, keeps them engaged, and encourages metacognition. These factors make this assessment a prime example of AISL.

Prior to 2012, the Advanced Placement Program (AP®) provided, in essence, three components to participating educators and students. The original component, administered by the College Board on behalf of colleges and universities nationwide since 1955, was the summative AP Exam, written and scored not by the students' own teachers but by committees of college professors and expert high school instructors. Second, there was a "Course Description" booklet, which contained a short outline of topics typically taught in the corresponding introductory college courses. Finally, the AP Program partnered with professional development centers to provide professional learning workshops, primarily focused on familiarizing teachers with exam details, scoring standards and rubrics, and techniques for teaching advanced topics.

In 2002, the National Research Council and the National Science Foundation issued *Learning and Understanding*, a report that indicated that the primary goal of AP and other advanced educational programs should be to help students develop a deep understanding of the organizing concepts and principles in all disciplines, and accordingly, curricula should focus on a reasonable number of concepts.[1]

In the decade that followed, College Board convened cognitive scientists and experts in each discipline, and from 2012, began implementing sweeping changes across the suite of 35 AP courses and exams, such that by Fall 2019, each AP course was redesigned, and anchored in a short list of transferable disciplinary skills that would now be the focus of each exam question. These skills became the spine of each AP course, recursively embedded within a finite body of content that would serve as a transparent compact with AP teachers about the full scope of content that could appear on an AP Exam.[2]

In short, this redesign of the 35 AP courses required a willingness for the sponsoring organization, College Board, to step away from an all-inclusive approach to course and exam topics—an approach that reflected the wide variation in content selected by the thousands of faculty and adjuncts who teach the college courses from which AP scores exempt students. Instead, the AP Program developed a transparent scope and sequence for each AP course, one informed as much by cognitive science researchers as by subject-matter experts in each field.

---

1 *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools,* Gollub, Jerry P., Bertenthal, Meryl W., Labov, Jay B., and Curtis, Philip C., Eds. National Academy Press, 2002.

2 Drew, Christopher, "Rethinking Advanced Placement," *New York Times,* January 11, 2011.

This change required delineating content formerly eligible for inclusion, as off-limits, and outside the scope of the AP Exam. Because there is no perfect consistency in the topics valued by the approximately 4,000 colleges and universities that utilized AP scores to place students out of introductory courses on their campuses, the AP Program incurred some degree of risk that these changes would alienate a subset of faculty whose favored topics were not included on the AP Exam. To minimize that risk, the AP Program conducted extensive analyses of syllabi from a range of institutions receiving AP scores, generated a comprehensive list of all topics appearing in college syllabi, and asked faculty to rate each topic's essentiality as a prerequisite for successful further study of the discipline on campus. Topics were then removed if they did not have high average ratings as essential foundational content. In parallel, the AP Program partnered with faculty and AP teachers to conduct exam timing analyses with the goal being to determine an appropriate amount of exam content ensuring adequate instructional focus on the recursive, transferable discipline skills.

Another significant improvement made possible because of the redesign of AP's focused course topic delineation is the design and delivery of free formative course assessments for all AP students. In the past, students and teachers had no way to check progress and calibrate learning and performance to an external benchmark, until they received their summative AP Exam scores each July—too late to make use of such information for that year's population of learners. The AP Program released AP Classroom at the start of the 2019 school year in conjunction with the 35 redesigned AP subjects. For each topic in every AP course, the AP Classroom platform provides daily instructional videos from a racially diverse group of expert AP teachers, daily formative practice questions teachers can assign before class, and an associated student data dashboard for instructors. The instructor dashboard provides teachers the opportunity to focus their instruction on correcting student learning misunderstandings, rather than dedicating precious instructional time to content or skills that students are already demonstrating well.

Accordingly, AP Classroom builds on the redesign of AP Exams to provide learners and their teachers with real-time feedback on topics they've mastered, skills they're developing, and how to focus further practice where the need is greatest. As a result, the usage levels are high, as is teacher satisfaction. In the 2022–23 academic year, the students taking AP classes watched a total of 66 million instructional videos and took 45 million formative assessments, generating an unprecedented amount of

direct and relevant instructional feedback for themselves and their AP teachers. Over 80% of AP teachers use AP Classroom resources, and nine out of ten teachers report it helps prepare for the exam and learn course content[3].

The free AP Classroom resources, anchored in the redesigned exams and course frameworks, now enable a cycle of teaching and learning supports that connect formative assessment data to instruction and learning, let alone preparation for the summative AP Exam, as Figure 1 depicts:

Figure 1.

1. **Plan**
   Unit Guides
   Professional Learning Videos
2. **Teach**
   AP Daily Videos
   Course-Specific Resources and Activities
3. **Practice**
   Topic Questions
   Question Bank
4. **Assess and Check for Understanding**
   Progress Checks
   Question Bank
5. **Get and Give Feedback**
   Individual Assignment Feedback
   All Assignment Report
   Progress Check Report
   Content & Skills Performance Report
6. **Review and Prepare**
   Question Bank
   Practice Exams
   AP Daily Review Videos

Incorporating Projects and Portfolios into AP Assessments

AP Exams have traditionally been defined by a single, three-hour examination at the end of a course, determining whether a student earns a qualifying score for college credit or placement. However, this approach has begun to shift with the introduction of performance tasks and portfolios as integral components of the AP assessment. As of 2024, these assessment models are employed in seven courses, covering approximately 400,000 exams. While AP Art and Design has utilized portfolio development for decades, this concept has more recently been adopted in

3 Aggregate of AP teacher surveys from Sep 2023-April 2024, n=2300

courses like AP Seminar and AP Computer Science Principles (CSP), among others. This shift reflects an organizational belief that performance tasks allow for deeper instruction and learning, provide more authentic assessments of skills, and make learning more engaging and relevant for students. AP's existing project-based assessment shows evidence of improved student performance and strong demand from both teachers and students for incorporating such projects into AP courses and exams. However, while projects have demonstrated their value and addressed community needs, challenges and open questions remain about how best to implement them in the AP Program.

Benefits of Performance Tasks

The introduction of performance tasks in AP assessments has yielded several promising outcomes. For one, students in these courses tend to have more success on the assessment, with students of similar levels of academic preparation being more likely to earn qualifying scores than in other AP courses and with high success rates among Black, Hispanic, and first-generation students. Additionally, courses with performance tasks like AP Seminar and AP CSP contribute to strong student performance not only on the AP Exams themselves but also in subsequent coursework and college. For example, students who take AP Seminar tend to earn higher first-year GPAs and have better retention rates in college than students who do not take APs[4]. Similarly, AP CSP often serves as the first AP STEM experience for many underrepresented students, and those who take it are more likely to pursue further studies in computer science and related fields[5].

Challenges and Open Questions

Despite these successes, challenges and open questions remain as AP continues to incorporate performance tasks into AP assessments. One of the main challenges is ensuring the security, validity, and consistency of these assessments—AP's core value proposition. Performance tasks, by their very nature, are more difficult to standardize than traditional exams. This challenge is compounded by the introduction of generative AI tools like ChatGPT, which raises new questions about ensuring authenticity.

---

4 Sanja Jagesic, Maureen Ewing, Jing Feng and Jeff Wyatt, "AP Capstone™ Participation, High School Learning, and College Outcomes: Early Evidence," College Board (2020).

5 Jeff Wyatt, Jing Feng, and Maureen Ewing, "AP Computer Science Principles and the STEM and Computer Science Pipelines," College Board (2020).

In the 2022–2023 school year, when tools like ChatGPT and DALL-E became widely available, AP initially attempted to enforce a ban on AI use in performance tasks. However, it became clear that this approach was neither practical nor beneficial. Instead, College Board is in the process of shifting over time to a policy of responsible integration, starting with AP Seminar and AP Computer Science Principles, allowing students to use AI tools in ways that support their learning while still ensuring that they demonstrate mastery of the material. Even with new policies in place, there are likely to be further issues to resolve as students and teachers learn more about both the benefits and shortcomings of generative AI tools.

Another challenge is the relatively low submission rates among Black and Hispanic students in courses that include performance tasks. Understanding the reasons behind these disparities is critical, as is finding ways to support all students in completing these tasks. This might involve rethinking guidance on how the tasks are administered, changing the performance task format or providing additional resources to help students succeed.

Building and scoring effective performance tasks and instructional resources requires significant resources and expertise. We are still in the process of developing archetypes for these tasks, balancing the need for valid assessment with the goal of providing space for student choice and creativity.

AP and Assessment in the Service of Learning

The literature on assessment in the service of learning highlights a number of different ways in which assessments can move beyond measurement and serve to enhance or improve learning. Whether it is by modeling expectations for test-takers, providing key insights to teachers, or establishing markers of progress, assessments can be used to improve learning outcomes. Of particular interest to assessment design is the way in which assessments can support the process of learning. This can be done by creating assessments that drive student motivation, engage students in some way, and promote metacognitive skills. In the areas of motivation and engagement, assessments can do things like providing "meaningful referents…that complement the previously existing cognitive frameworks of the student" (Qualls, 1998, p. 298). When students recognize themselves in the material presented on the assessment, they are more likely to be engaged and motivated to perform well. Encouraging metacognition, or a reflection on one's own thinking, means creating an assessment that encourages test-takers' "monitoring their own

understanding, predicting their performance, deciding what else they need to know, organizing and reorganizing ideas…[to] help them advance their understanding" (Earl, 2006, p. 4).

With its focus on providing support materials for teachers and students that allow opportunities for real time feedback, the redesign of AP courses that was completed in 2019 solidified the AP Program's commitment to the ideals of Assessment in the Service of Learning. And the shift toward performance tasks moves the needle even further. Of course, AP has had a model for this approach to assessment since 1972. AP Art and Design as a program has always modeled these process-related aspects of assessment in the service of learning, and with its own redesign, it now has additional factors that can motivate test-takers, support engagement, and encourage metacognition. Before we examine how these ideas play out in the redesigned course and portfolio assessment, however, we should provide some basic details about AP Art and Design, its history and its current program structure.

Art and Design: Pioneering Project Based Assessment in AP

In 1972, College Board pioneered standardized student portfolio submissions and assessment through AP Studio Art. Since then, participating high school students have had the opportunity to gain college credit or advanced placement in drawing, 2-D design, and 3-D design by achieving a passing score of 3 or above (on a scale of 1–5). As part of the AP Program's intentional course redesign focus, AP Studio Art was reimagined and became AP Art and Design in 2019. The revised course includes an increased focus on student inquiry to guide art-making through the Sustained Investigation portfolio component. In the Sustained Investigation, students answer two writing prompts:

1. Identify the inquiry that guided your sustained investigation.

2. Describe ways your sustained investigation developed through practice, experimentation, and revision.

The 2023 AP Art and Design *Course and Exam Description* (CED) defines a sustained investigation as "an inquiry-based and in-depth study of materials, processes, and ideas over time" (p. 43). In this portfolio component, students are encouraged to discover, explore, question, reimagine, practice, experiment, and

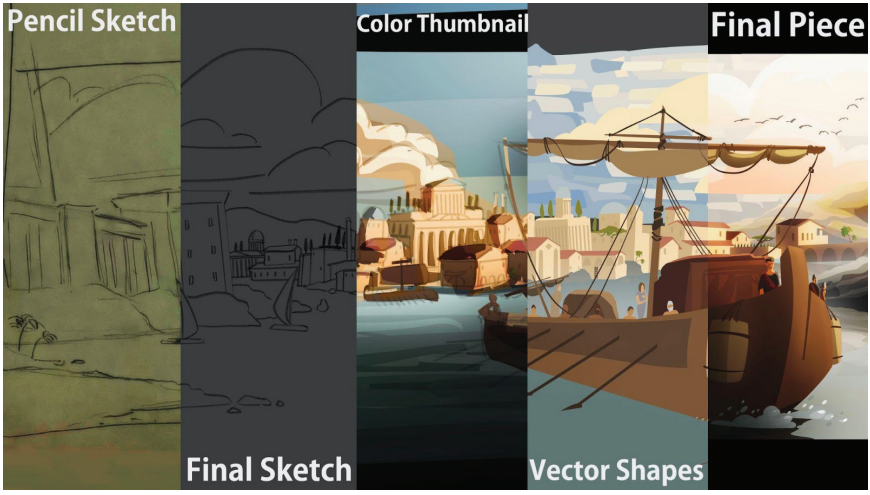
revise to demonstrate synthesis of materials, processes, and ideas. Students develop their inquiry based on personal experiences to create unique and original artworks. "Experiences can be documented by recording observations and perceptions related to an experience" (p.14) using "any materials, processes, and ideas as long as the work is the student's original creation" (p. 35). Thus, students are free to choose ideas, materials, and processes that are the most meaningful and personal to them (Escoffery et al., 2025). During the annual AP Art and Design exam assessment, readers (raters) often note that the most exciting and engaging portfolios to score are those derived from student passions, personal lives, and their art-making discoveries.

When assessing the sustained investigation portfolio component, readers use an analytic rubric to measure four art-making practices (See Appendix A):

1. guiding inquiry,
2. practice, experimentation, and revision in art-making,
3. synthesis of materials, processes, and ideas in art-making, and
4. portfolio skills.

Each Sustained Investigation analytic rubric row contains decision rules defining how a rater can apply a score of 1–3 to best award student achievement. In this portfolio component (worth 60% of the overall exam score), students demonstrate their thinking through art-making in writing and digitally submitted images and works (e.g., sketchbook pages, mood boards, mindmaps, experimental or process images, and final artworks). For example, in Figure 2, Daniel Stordahl, whose portfolio was featured in the 2024 AP Art and Design Exhibit (Stordahl, 2025), shares a digital image composite demonstrating the drawing process he used to tell the "story of young Julius Caesar's capture by pirates in 75 BC and his vow to return and destroy them" (para. 2). The written evidence accompanying his process work elucidates material choices and conceptual and physical process(es). Daniel describes his materials as "Paper, pencil, Adobe Fresco, iPad" while his processes include "Compose sketches, plan color/light, block shapes in vectors, render shadows/gradients, cinema border" (Stordahl, 2025, para. 2). By including part of his finished artwork in this process work, we understand the progression and choices made from inception to completion.

Figure 2.



Note. From *Caesar Departs from Rome,* by D. Stordahl (2025), *2024 AP Art and Design Exhibit* (https://apartanddesign.collegeboard.org/2024-student07).

In this image, the process writing informs the viewer's interpretation and when paired with Stordahl's inquiry statement (written evidence), the investigation into the relationship of exploring *The Revenge of Julius Caesar* through cinematic techniques to tell a story and convey emotion in a single shot is evident:

> *Throughout every step, I was intentional about contributing to the bigger picture of the story. For example, in "Caesar Departs from Rome," I wanted viewers to feel the power and glory of Rome, represented by the sunlit city in the background. At the same time, I positioned Julius Caesar venturing toward a cloud-covered area, symbolizing the danger and uncertainty of the outside world while foreshadowing the peril he would encounter. This deliberate process ensured that each element added meaning and contributed to the narrative.* (Stordahl, 2025, para. 6)

It is important to note the assessment requires that inquiry guides art-making. Thus, the written inquiry statement is a valuable tool aiding the student's ability to narrow their art-making exploration and discovery to a targeted focus and clarifies the presented visual evidence. However, at the heart of the sustained investigation portfolio component is an art-making focus on practice, experimentation, and revision of materials, processes, and ideas. The written inquiry statement guides art-making exploration and substantiates the visual images submitted for evaluation.

The sustained investigation process aligns closely with PBLWorks's Gold Standard PBL: Essential Project Design Elements, which emphasizes sustained inquiry, student voice and choice, opportunities for critique and revision, and reflection (Buck Institute for Education: PBL Works [PBL Works], n.d.). In AP Art and Design, students are given the time and space to work like artists, gradually developing a portfolio that reflects both their skills and their creative process. The rubric for the sustained investigation portfolios emphasizes inquiry and student reflection, promoting a cycle of learning, reflection, and revision. Students are encouraged to describe how their sustained investigation was guided by inquiry and demonstrates practice, experimentation and revision of materials, processes, and ideas.

In contrast to the Sustained Investigation analytic rubric, readers use a holistic rubric to assess the second portfolio component, Selected Works (Appendix B). This component measures student accomplishment in portfolio skills and their ability to synthesize materials, processes, and ideas in finished artworks. Although the Selected Works component does not include formal writing prompts, student writing accompanies each final work, providing information on the students' idea(s), materials, and process(es). In Figure 3, 2023 AP 3-D Art and Design student Audrey Nordfelt created a composite image showcasing scale and detail in her sculptural work. Nordfelt provides information on her idea of perception and developing individual meanings. The idea explanation aids interpretation and understanding of the visual image, and when combined with a materials description "cone 10 clay, high fire glazes layered for custom effect, K9 & Las Vegas Red" and process(es) "sculpted hollow form, added hollow tentacles, factoring in balance, fired in reduction, added base" (Nordfelt, 2023, para 1), the viewer gains insight into how the artwork was developed and executed to fulfill the student's vision.

Figure 3.

Note. From *Currents,* by A. Nordfelt (2023), *2023 AP Art and Design Exhibit* (https://apartanddesign.collegeboard.org/2023-student05).

When scoring the Selected Works portfolio component, readers review the digitally submitted student works, the accompanying text clarifying idea(s), materials, and process(es) and use a scale of 1–5 to assign a score (See the Selected Works Scoring Guidelines in Appendix B). The Selected Works are worth 40% of the student's total AP Art and Design score.

Section II: Training and Course supports

AP Art and Design Course Rubrics and Scoring Guidelines

The rubrics and scoring rules (Appendixes A and B) are consistent from year to year and available for teacher and student use on College Board's website, AP Central. AP Central also hosts a web page for each portfolio (2-D Art and Design, 3-D Art and Design, and Drawing) and includes sample student portfolios, providing written and visual evidence for each rubric score point in the sustained investigation and selected works portfolio components. The samples include written commentary from experienced AP Art and Design readers (comprised of high school art educators and higher education faculty) who relate student work to the rubric and describe how each sample achieved a score point. Many high school teachers use the sample student portfolios in conjunction with the course rubrics in low-stakes formative assessments as a way for students to discuss, critique, and practice applying the course rubrics to visual and written evidence. For example, students may use the rubrics to guide conversation during an in-class critique of student artwork and writing. Students might also work in small groups to discuss art-making progress using specific course rubric content (including definitions) as a focus. In gallery walks (where all student artwork is on display for review), students can use Post-It notes and write feedback aligning with rubric language. The AP Art and Design course rubrics are often printed and added to student sketchbooks for ongoing personal review and reference.

Most importantly, the rubrics direct students toward essential art-making practices inherent to learning and growing through practice, experimentation, and revision of materials, processes, and ideas through an inquiry-based approach. For example, to achieve the highest score in Row B of the Sustained Investigation rubric, students must provide "*visual* evidence of practice, experimentation, and revision demonstrat[ing] development of the sustained investigation." This statement ensures students provide evidence that they have practiced, experimented, and revised their materials, processes, and ideas in pursuit of an in-depth investigation over time. To achieve the highest score in Row C of the Sustained Investigation, students must provide evidence of the "visual relationships among materials, processes, and ideas and demonstrate synthesis." Students who achieve synthesis have practiced, experimented, and revised throughout their portfolio as they worked towards the coalescence of materials, processes, and ideas. The rubric structure outlines ways students can successfully produce art while providing language that guides discussion and feedback through various formative assessment practices.

Teaching and Learning Supports

The AP Art and Design *Course and Exam Description* (CED) is a conceptual framework outlining course skills and content applicable to lesson planning through Big Ideas, Learning Objectives, and Essential Knowledge Statements. It is available for download from each AP Art and Design portfolio (2-D, 3-D, or Drawing) hosted on College Board's webpage, and instructors are encouraged to print and share the CED with students. As noted above, the redesign of AP Programs allowed College Board to create new resources for teachers and students. For AP Art and Design, these resources serve to clarify CED expectations. Experienced high school AP Art and Design teachers and college or university faculty host a series of on-demand short videos, called AP Daily Videos, in College Board's learning management system, AP Classroom. The AP Daily Videos clarify ideas in the CED by offering targeted lessons that teach curricular concepts in 7–15 minute segments. AP Art and Design teachers can assign videos to their students through AP Classroom to watch as part of daily work, and students can review as a class, as a small group, or individually. To ensure students understand how the AP Art and Design rubrics are applied when scoring student portfolios, AP Classroom additionally hosts rubric training videos that compare student examples to course rubrics and explain how students achieve rubric points. The same rubric training videos are used to norm readers to the exam requirements and rubric application during the annual AP Art and Design portfolio assessment (Reading). This transparency ensures all students and teachers can access the current visual and written rubric explanations as student work develops and before the final portfolio assessment occurs. Using the CED and companion AP Daily Videos, students are guided on developing their visual art images and works to align with the summative course rubrics and scoring guidelines.

Finally, College Board's website page for AP Arts Webinars also hosts free, on-demand webinars such as *Best Practices on Using the AP Art and Design Rubrics*. These resources are designed to be flexible, allowing teachers and their students to watch AP Classroom and webinar videos together, individually, or in small groups. This adaptability enhances any school's AP Art and Design curriculum, catering to different learning styles and situations.

AP Art and Design Exhibits

College Board's AP Art and Design Exhibit (College Board, 2024) is an annual exhibition showcasing exemplary student artwork. During the yearly AP Art and Design Reading, leaders review student portfolios and choose student artworks representing diverse artistic approaches and ideas, student demographics, and school locations. In total, each year's exhibit includes an average of 50 students. After the initial curation process, students are invited to submit high-resolution images of their selected artworks and create a student statement for publication. In the statements, students respond to prompts that guide them in explaining how they came up with their inquiry idea and how it developed during the school year. Their explanations also clarify the portfolio rubric (e.g., their intentionality in choosing materials and developing processes to support ideas and achieve synthesis). The guided student explanations showcase their work and teach other students (and teachers). Additionally, the student's art teachers and school leaders share best practices for ways in which they support teaching and learning in AP Art and Design. Teachers often write about how they support inquiry-based learning, and school leaders explain how they support and promote the visual art program in their schools. The power of the AP Art and Design Exhibit's design is to intentionally showcase student artwork and serve as a teaching and learning tool through exemplars and detailed student, teacher, and school leader best practice explanations. The exhibit is available on the internet and linked to AP Classroom so that teachers can refer to it as a resource that supports the AP Art and Design rubric and scoring guidelines. The exhibit has over 230,000 visits annually, making this teaching tool a valuable resource for instruction and assessment.

Section III: Theories of Assessment in the Service of Learning

All of these elements of the AP Art and Design portfolio program combine to make it an excellent example of Assessment in the Service of Learning. Edmund Gordon (2020) defines this idea as "an approach to Pedagogy in which assessment, teaching, and learning are organically interrelated such that these three processes are dialectically and reciprocally employed each in the service of the other" (p. 73). In many ways, AP Art and Design serves as a perfect illustration of how assessment can be "organically interrelated" with teaching and learning. While it is ultimately a summative assessment leading to the awarding of a final AP score that can be used by colleges and universities to grant students placement or credit for

work done in high school, the AP Art and Design portfolio is intentionally process-focused in a way that allows it to work in the service of learning. As Gordon (2020) puts it, "Assessments should be designed so that the processes of student thought and creation are visible…Portfolios can make visible the scaffolding, both from the teacher and the students' own processes that resulted in the product" (p. 74). The AP Art and Design portfolio works in just that way; in fact, the portfolio submission requires that students detail the processes of thinking that led to the works they have created.

As we have seen, the design of the portfolio requirements and of the rubrics used to evaluate those portfolios specifically creates opportunities for students to demonstrate the processes they used and to explore the steps of learning that took place over the course of the development of their portfolios. In the Sustained Investigation section of the portfolio, students can submit images that document their art-making process, and the rubrics specifically ask for evidence of practice, experimentation, and revision (See Appendix A). The written evidence that students supply also provides opportunities for them to reflect on and discuss their process, the decisions they made, and any development or revisions that came from their investigation of the inquiry topic. Thus, in Row B of the Sustained Investigation rubric, raters are asked to look at both the visual and written evidence. To achieve the highest score on that Row, the work must demonstrate the following: "*Visual* evidence of practice, experimentation, and revision **demonstrates development** of the sustained investigation. **AND** *Written* evidence describes ways the sustained investigation **developed** through practice, experimentation, and revision" (See Appendix A). In the literature on Assessment in the Service of Learning, a focus on *process* is discussed as a multifaceted aspect of assessment that pushes one beyond the realm of mere measurement and into the service of learning.

One aspect of *process* that the AP Art and Design portfolio highlights is student motivation. In asking students to follow a line of inquiry through practice, experimentation, and revision, the portfolio demands and hopefully encourages students to exhibit a certain amount of motivation as they solve problems and develop artworks of their choosing. An inquiry driven curriculum places students in the driver's seat of their learning, providing autonomy (a basic tenet of motivation). In her discussion (2006) of assessment as a "powerful lever for learning," Lorna Earl notes, "In the medium and long term, assessment [holds] the

possibility of…influencing students' motivation as learners and their perceptions of their capabilities" (p. 4). Learning is not a static state that can be simply identified by an assessment, something a student has or has not acquired; rather, it is a "dynamic process" (Earl, 2006, p. 6) that requires active engagement on the part of the learner.

To keep students actively engaged in the learning process throughout the entire portfolio development process, the AP Art and Design program includes a number of features designed to increase student motivation. The heart of the portfolio is the Sustained Investigation section, and a quick look at the Scoring Criteria (See Appendix A) for this section demonstrates factors that are linked to motivation. First, the Sustained Investigation section is meant to be guided by an inquiry that the test-taker chooses based on their own specific interests, and the first row of the scoring rubric assesses whether or not there is an inquiry and to what extent that inquiry has guided the investigation. That is, test-takers are not simply asked to create individual works of art. They are asked to use their artwork as a vehicle to investigate and explore ideas that are of interest to them.

Additionally, students are asked to solve problems that arise as they conduct their investigation, and problem-solving is a key feature of motivation. As Earl puts it, "Not only are humans able to search for problems to solve; they appear to enjoy it" (2006, p. 5). In the case of AP Art and Design, students can encounter any number of interesting problems, from difficulties composing works to challenges using specific media. The course and portfolio are designed to encourage students to engage with those problems and learn from them. We can see this emphasis on problem-solving in the part of the rubric related expressly to whether or not the test-takers have engaged in "practice, experimentation, and revision [that] demonstrates development of the sustained investigation" (College Board, 2023, p. 41). Test takers can show evidence of this practice, experimentation, and revision by including process documentation in their portfolio. That could be a preliminary sketch or model that led to a more finished work, an image of a piece that was unsuccessful but provided a key idea, or documentation of an artistic idea as shown in Figure 3. Because the focus of the portfolio is on inquiry, investigation, and exploration, test-takers are not required or expected to include only polished, 'perfect' works of art. Because perfectionism can have a negative impact on student motivation (Fletcher & Neumeister, 2012), it can increase motivation to allow test-takers to include works that show growth or learning,

such as process pieces and works that were revised. The explicit focus in the portfolio requirements on "practice, experimentation, and revision" (See Appendix A) allows students the freedom to try new things and fail. In fact, a student's failures can increase motivation in a situation like this because "perfection" is not an expectation or a requirement.

Motivation, in fact, is a key factor enabling someone to continue when a task or process is difficult, and creating works of art can present difficult challenges. Such challenges for AP Art and Design students might be related to the use of a particular medium (paint, ceramics, digital photography, etc.), the attempt to find the proper style to use to communicate a given idea, or solving a problem related to composition, which was a struggle encountered in 2023 by Aanje Greymountain (Greymountain, 2023). In one artwork (See Figure 4), Greymountain was attempting to depict the ending of the Navajo story of the Hero Twins, the moment when the Twins bring the head of the evil giant back to their mother and grandparents. However, she struggled to find the right composition for the piece, something that would depict both the Twins and the head of the giant. As she notes, "I had a tough time creating this piece. For the life of me, I could not find out how to fit in the head of the giant despite it being the central element of the storytelling" and took "much trial and error" (Greymountain, 2023, para. 3) to get to the composition she ended up using, which we see in Figure 4. As we will discuss later, this kind of self-reflection or metacognition is further demonstration of the way that AP Art and Design operates as assessment in the service of learning, literally helping shape the artist's practice and process.

Figure 4.

Note. From *Hero Twins,* by A. Greymountain (2023), *2023 AP Art and Design Exhibit* (https://apartanddesign.collegeboard.org/2023-student01).

Course instructors for AP Art and Design can help students understand the importance of maintaining motivation in the face of interesting challenges. Greymountain's teacher, Greg Stevens, notes how the structure of the AP course and assessment served as effective motivation, saying,

> *Through practice, experimentation, and revision, Aanje successfully fulfilled her vision. Sketchbooks were filled with different compositions, details, and subject matter. What started as a verbal story was written down and divided into visual pieces. Those pieces were then vetted through critiques, self-analyzation, and cohesion. Nothing was considered sacred, and everything was up for discussion, debate, and revision… The College Board has provided a structure that allows students to make their art more authentic, conceptual, and personally fulfilling. It's not so much teaching the technical aspects but the behavioral traits of an artist* (Stevens, 2023, para. 2).

The AP Art and Design portfolio, then, is structured in such a way as to provide motivation for test-takers, helping them learn the "behavioral traits of an artist" (Stevens, 2023, para. 3) and giving them the tools to solve problems in ways that enhance learning. In addition, many of the tools discussed earlier in the article (e.g., AP Classroom videos and the Exhibition) give teachers the resources they need to help students maintain motivation while solving problems.

In the literature related to assessment in the service of learning, researchers note the importance of engagement as a major factor encouraging student learning, and engagement is another aspect of *process* that the Art and Design portfolio encourages. Dylan Wiliam, speaking of the forces that drive successful learning, notes, "[T]here is now a strong body of theoretical and empirical work that suggests that integrating assessment with instruction may well have unprecedented power to increase student engagement and to improve learning outcomes" (2011, p. 22). For AP Art and Design, the focus on inquiry, experimentation, and exploration in the portfolio requirements and in the evaluation criteria is designed to enhance student engagement in ways that the former AP Studio Art course and portfolio allowed but did not explicitly encourage. Although AP Studio Art originally had a Concentration section that allowed test-takers to focus on an idea of interest, certain aspects of the evaluation criteria rewarded mastery of technical skill over inquiry. For example, in the previous course, one of the bullets in the scoring guidelines describing the highest score point for the Concentration section read, "In general, the work is technically excellent" (College Board, 2019, p.7). And for many years, the Selected Works section of the portfolio was known as the Quality section, a name that emphasized the focus on mastery of technique and the creation of highly polished, finished works of art. AP Art and Design shifted its focus to inquiry-driven investigation. As part of the redesign process described at the beginning

of the chapter, College Board held extensive discussions with college professors and those who run foundation art programs at the college level. The predominant feedback was that college foundations courses prioritize inquiry and investigation over the creation of finished artworks. A quick look at the terminology defined in the scoring criteria reveals the redesigned AP Art and Design course does value exactly these inquiry-related concepts—development, discovery, experimentation, exploration, practice, process, and revision (College Board, 2023b). The glossary defines the key concept, inquiry, as "the intentional process of questioning to guide exploration and discovery over time" (p. 43). And this vision of inquiry, the call for students to ask questions and explore topics of interest to them, helps keep them engaged as they develop the works that are included in their portfolios.

Because students in AP Art and Design are exploring topics of interest to them, they are more likely to be engaged with their work, which leads to greater satisfaction. According to Naomi Holmes (2017), "Student engagement is intrinsically linked to two important metrics in learning: student satisfaction and the quality of the student experience" (p. 23). This sense of satisfaction can lead to enhanced effort and ultimately to stronger performance. There are many examples of engagement in successful portfolios submitted for AP Art and Design. For instance, Audrey Nordfelt, who took the AP 3D Art and Design course in 2023, started out feeling like ceramics were, as her teacher put it, "outside her comfort zone" (Frampton, 2023). As she worked on pieces for her portfolio, Nordfelt (See Figure 5) became increasingly engaged by the idea of perception because people kept telling her what they thought her artworks represented. As she says, "So many people would ask what I was making, and then they would tell me what they thought it was. For the most part, people saw it as different things. This made me curious about perception again. I decided to look into it and research human brains and how we process things we see. I learned that there are different steps to perception" (Nordfelt, 2023). Because she was engaged with this particular idea, Nordfelt was able to overcome her discomfort with the medium she was exploring and create work that was both meaningful to her and successful according to the portfolio scoring guidelines. For her work, *Currents*, featured in the 2023 AP Art and Design Exhibit (College Board, 2023a), she noted that different people saw different shapes or creatures (e.g., anemone or octopus) in it.

Figure 5.

Note. From *Currents*, by A. Nordfelt (2023), *2023 AP Art and Design Exhibit* (https://apartanddesign.collegeboard.org/2023-student05).

Nordfelt's research on perception, or the notion that "because we all have learned different things and lived different lives, we all have different knowledge and use that knowledge to perceive things we see differently" (Nordfelt, 2023), promoted the kind of engagement required to create a successful portfolio. As her teacher notes, "Sometimes, students must be encouraged to keep going even when they do not know how it will happen. She went on to win the best of show in our district art competition and created an amazing AP Art and Design portfolio" (Frampton, 2023, para 6). In this case, engagement played a large part in helping this student "keep going."

Section IV: Formative Feedback cycles Supporting Metacognition

In the case of AP Art and Design, the final, summative assessment (the portfolio) is designed to encourage student attention to process, causing the portfolio to function in many ways like a formative assessment. Students put together their portfolios over the course of a year or longer, with regular opportunities for teacher feedback to guide student revisions leading to changes in subsequent works that make the final portfolio more successful. Building on Arkalgud Ramaprasad's classic definition of feedback (1983), D. Royce Sadler notes that "information about the gap between actual and reference levels is considered as feedback only when it is used to alter the gap" (1989, p. 121). That is, the feedback that teachers provide on AP Art and Design portfolio work can be used to improve performance. Thus, it meets William's (2011) requirement of being "information generated within a particular system, for a particular purpose" (p.3), rather than information "separated…from its instructional consequences" (William, 2011, p. 3). Within the AP Art and Design classroom, teachers are consistently working with students to revise and refine works, explore new ideas that could further the inquiry, and learn from both mistakes and successes. The scoring guidelines, which give points for successful experimentation and revision, are explicitly constructed to reward exactly this kind of formative feedback.

Furthermore, the emphasis within the portfolio requirements and the scoring criteria on inquiry keep the students actively engaged in the learning process. The example works discussed above show how the program is designed to encourage motivation and engagement by having students pursue a line of inquiry that is interesting to them (a traditional story important to the student's culture or an intellectual idea that the student finds fascinating). As Earl (2006) notes,

"Learning was long thought to be an accumulation of atomized bits of knowledge that are sequenced, hierarchical, and need to be explicitly taught and reinforced. Learning is now viewed as a process of constructing understanding by attempting to connect new information to what is already known so that ideas have some personal coherence" (p. 4). Following a line of inquiry through experimentation and revision in the AP Art and Design portfolio requires students to do exactly that—construct understanding by connecting new information to what is already known.

The formative feedback cycles that the AP Art and Design portfolio allows for, and that the scoring criteria encourage, support students in a specific form of feedback, namely metacognition, which "occurs when students personally monitor what they are learning and use the feedback from this monitoring to make adjustments, adaptations and even major changes in what they understand" (Earl, 2006, p. 7). Take, for instance, the focus in the Sustained Investigation section on revision as one of the key skills test-takers need to demonstrate. Throughout the process of developing a portfolio, a student is asked to look at the work they have already created and make adjustments based on what they have learned, what has worked, and what has not come across as they expected. That is, the student must engage in metacognition in relation to the works that have already been created, thinking about the thinking that went into each piece and making adjustments as they progress. Using this type of metacognition can help learners to "understand and control their own cognitive processes" (Hands & Limniou, 2023, p. 125). Student development of metacognitive strategies has been tied to better learning outcomes, such as moving from surface to deep learning approaches (Hands and Limniou, 2023), and it is theorized to "play a fundamental role in guiding students' learning across domains" (Taouki, Lallier, & Soto, 2022, p. 921). Supporting these metacognitive activities was an active goal of the Art and Design redesign process, and we see clear evidence that the new portfolio requirements do, indeed, encourage this kind of thinking.

As the AP Art and Design *Course and Exam Description* points out, the process of investigation that is at the core of the work done to develop a portfolio "can confirm and challenge thinking, revealing connections and opportunities" (College Board, 2023, p. 14). Students are encouraged to focus on this metacognitive process both by the portfolio design with its emphasis on inquiry and by the fact that Row B on the Sustained Investigation scoring guide explicitly assesses whether the works demonstrate Practice, Experimentation, and Revision (See Appendix A). That is,

students are directly rewarded for metacognitive practices like making revisions based on examining and thinking about the results of an earlier attempt.

For an excellent example of the way metacognition can influence the development of the artworks going into a specific AP Art and Design portfolio, we can look at the work of Elizabeth Tian (See Figure 6), who submitted a Drawing portfolio in 2023 and had work that appeared in the 2023 Exhibition (College Board, 2023). According to Tian, "The state of mind can be a place of disruptions, brawls, celebrations, or serenity" (2023, para. 2). Because she was aware of and able to reflect on those different, conflicting states of her mind, she was determined to create works that "depict a visual strain that reflects one's emotional strain" (Tian, 2023, para. 2) related to the pressures that society places on each individual due to unrealistic expectations. In the piece *Gasping*, we see this metacognitive exploration developed visually. Tian claims this piece explores the "accumulation of immense pressure that is overwhelmed by its constantly changing surroundings" (Tian, 2023, para. 9). In the work, Tian includes "cheeky laughing and screaming mouths, frantic eyeballs, and crooked, yellowed teeth" to visually demonstrate the idea that "society tries to draw people into what they see, say, and feel" (Tian, 2023, para. 3). Thus, her thinking about the way society impacts a person, creating tension and distortions, led to Tian's experimentation with both content (exaggerated and distorted features) and form. *Grasping* is a self-portrait, in which the artist is surrounded in a swirl of grotesque figures, representing directly the kind of social pressure Tian is investigating, depicting her "struggle to cry out, gasping for relief" (Tian, 2023, para. 3). And yet, the work contains balance and symmetry. There is order and beauty that indicates the relief and peace that lie beyond the tension Tian is exploring. As she notes, the tension we all experience "will soon be released because we evolve as we experience it" (Tian, 2023, para. 2). It is this level of metacognition and recognition that the AP Art and Design program both allows and encourages students to reach.

Figure 6.

Note. From *Flooding,* by E. Tian (2023), *2023 AP Art and Design Exhibit* (https://apartanddesign.collegeboard.org/2023-student14).

Section V: Engaging Community through AP Art and Design: Learners, Teachers, Administrators, and Families

Presentation

When students embark on their journey to produce a portfolio of work for AP Art and Design, the production and exhibition of their work is often a community affair. From informal class critiques to formal end-of-year art shows, the visual art students present their work throughout the art-making process. College Board's CED (2023b) speaks to presentation and audience engagement like Nordfelt (2023) sought to engage others through perception. Both focus on interpretation as part of presentation. Essential Knowledge Statement 3.F.1 informs teachers and students that

> *Presenting works of art and design to viewers for interpretation involves making decisions about what to show, when to show it, how to show it, and to whom it is shown. Different ways of presenting work can lead to different interpretations—even for the artist or designer who made the work. The artist or designer has the power to affect how materials, processes, and ideas within a work are perceived, based on decisions they make about how they present or display the work* (p. 27).

Students are thus directed to intentionally engage their audience through presentation choices to affect interpretation. Artworks, by nature, are meant to be viewed and interpreted, leading to conversation and dialogue about artistic intention and purpose. The CED further directs student artists that presentation "can include communication[s] between the artist or designer and the viewer" to "inform thinking and making" (College Board, 2023b, p. 27). Communication may occur through discussion, writing, and even visual responses. The CED advises students to consider how "documentation can include viewer interpretations of the work presented. Documentation of presentation becomes a resource for the artist/designer and it can be shared with viewers" (College Board, 2023b, p. 27). A student artist has the capacity to engage others by developing a dialogue through presentation processes. In part, this kind of dialogue begins in the classroom through critiques focused on presenting, interpreting, and providing feedback on artwork.

Formative assessment

Visual art critiques are an integral form of formative assessment in an art and design curriculum. By their nature, art class critiques develop a sense of peer community through shared purpose and meaningful engagement around art-making. Students struggle together to communicate ideas, improve art-making practices, discuss processes, and create a finished project. Art critiques can be short teacher and student feedback sessions lasting minutes to several days of classroom conversation focused on an entire class's artwork development. An art critique typically involves a presentation of the final artwork or work in progress followed by a discussion of the ideas, material choices, and processes used. Some critiques are teacher-directed, while others are collaborative activities with whole class engagement. Dan Kuffel, a teacher whose student College Board curated into the 2023 AP Art and Design exhibit, wrote in his teacher statement that he supports student learning by

> *encourage[ing] students to work in small groups to promote the cross-pollination of ideas. Having a sounding board, opposing perspective, friendly ear, or complete collaborator as you create your best work. These also act as informal critiques while the works are developing. Work is shared and refined, and usually, your friends will tell you the truth* (Kuffel, 2023, para. 7).

In this capacity, learners are engaged and motivated to participate, to help and encourage each other, to develop friendships, and to understand ways to improve. Honest communication and relationship-building through art production are foundational to building trust. Making and presenting art is a vulnerable process, and trust is integral to supporting authentic communication and creativity. Maggie Jones, another teacher whose student, Aundrea McCarthy, was curated into the 2023 AP Art and Design Exhibit, wrote in her teacher statement that "critique sessions serve as collaborative forums, where students offer each other constructive feedback, fostering a sense of community within our creative space" (College Board, 2023, para. 3). Through critique, learners present and reflect, document their learnings from peer or teacher reviews, and discuss learnings in ways unique to the visual arts. Critiques provide regular feedback cycles that enhance student ideas, skills, and artistic growth. Presentation and critique practices build community engagement in the classroom that can parallel how professional artists engage with others.

School and Community Engagement

College Board's CED guides teachers and school administrators to consider that "students need time and resources to engage with art and design in the classroom, school, and in the local community as well as in museums and galleries (in person and virtually)" (College Board, 2023b, p. 5). While visual art critiques are one way to engage and build community, extending the visual arts curriculum into the community is another. It is crucial for students to engage with art and design in various settings, as it broadens their perspective and enhances their learning experience. Virtual or in-person field trips or gallery visits allow students to engage with how adult artists think, create, and present their art. When students perceive how adult artists grapple with art-making to communicate ideas with their processes, they may gain insight into how they, as young artists, may fit into a broader art-making community. This kind of external connection or meaning-making builds purpose and reinforces internal motivation. When AP Art and Design students discover how or why their voice matters in a larger context, their inquiry goals become even more meaningful.

Visual art teachers are generally creative in forming community within their classroom, school, and local community. In a statement about her featured student in the 2023 AP Art and Design Exhibit, educator Emily Lemp writes, "...we often have showcases and gallery walks throughout the school year and partnerships with outside organizations that host some of these events, such as the law firm Cleary Gottlieb" (College Board, 2023, para. 2). By working with an external sponsor, Lemp can build relationships between a law firm and student artists to support and engage the school community in ways appropriate to the school context. Another featured educator in the 2023 exhibit, Suzanne Zimmerman, writes in her teacher statement that the AP Art and Design program advances students because she builds relationships with local designers and creative mentors. She adds,

> *We show work annually at our local art center in a professional gallery, a collaboration that has led to internships and employment through networking in the community. Critiques, competitions, school art shows, and creating work to sell for charity cultivate a comprehensive picture of how to be an engaged artist and activist in practice. We try to help our young artists thrive professionally and personally by learning to apply creative problem-solving, community, confidence, and perseverance in the art studio and their other life adventures* (Zimmerman, 2023, para. 3).

Building community partnerships between students, the school community at large, and the surrounding community, including businesses and art studios, creates rich programming to engage and support students.

Section VI: Conclusion

AP Art and Design, then, offers an excellent model demonstrating the different process-related aspects of Assessment in the Service of Learning. Because it allows opportunities for formative assessment throughout the development of the portfolio, the program engages students (and their broader communities both inside and outside of the classroom), enhances motivation, and fosters the use of metacognitive strategies. The *Course and Exam Description* (College Board, 2023a) explicitly drives students toward these different aspects of learning, as it puts the focus of the course and the exam on investigation, experimentation, and revision. Those ideas are essential to the development of the work, but also to the thinking that goes into all aspects of the course, from the decision to explore a particular Sustained Investigation topic to the discussion of works with classmates and the broader community through to the selection of works to present in the portfolio. Each step in the process allows the students to examine the decisions they have made, look at the impacts of those decisions, and adjust. This metacognitive work fosters deeper learning (Hands & Limniou, 2023), which is evident in the outcomes seen for students who have taken AP Art and Design. More studies should be done to evaluate the impact of AP Art and Design on student learning, but the preliminary results (Escoffery et al., 2025) point to the idea that theories of assessment in the service of learning do result in strong learning gains.

References

Buck Institute for Education (PBLWorks). (n.d.). *Gold standard PBL: essential project design elements*. PBL Works. https://www.pblworks.org/what-is-pbl/gold-standard-project-design

College Board. (2019). *AP Studio Art scoring guidelines* (Effective 3/15/2019) [AP Reading Training Material].

College Board. (2023a). *AP Art and Design course and exam description* (Effective Fall 2023). https://apcentral.collegeboard.org/media/pdf/ap-art-and-design-course-and-exam-description.pdf

College Board. (2023b). *2023 AP Art and Design Exhibit*. https://apartanddesign.collegeboard.org/2024-ap-art-and-design-exhibit

College Board. (2024). *2024 AP Art and Design Exhibit*. https://apartanddesign.collegeboard.org/2024-ap-art-and-design-exhibit

Council, N. R., Division of Behavioral and Social Sciences and Education, Education, C. F., & Committee on Programs for Advanced Study of Mathematics and Science in American High Schools. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools* (New ed.). National Academies Press.

Drew, C. (2011, January 9). Rethinking Advanced Placement. *New York Times*. https://www.nytimes.com/2011/01/09/education/edlife/09ap-t.html

Earl, L. (2006). Assessment—A powerful lever for learning. *Brock Educational Journal, 16*(1), 1–15.

Escoffery, D. S., Fletcher, K. E., & Stone-Danahy, R. A. (2025). "A search for my voice": Socioculturally responsive assessment in AP Art and Design. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 262–282). Routledge. https://doi.org/10.4324/9781003435105

Fletcher, K., & Neumeister, K. (2012). Research on perfectionism and achievement motivation: Implications for gifted students. *Psychology in the Schools, 49*(7). 668–677.

Frampton, R. (2023, December 1). *Teacher statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student05

Gordon, E. (2020). Toward assessment in the service of learning. *Educational Measurement: Issue and Practice, 39*(3), 1–142.

Greymountain, A. (2023, December 1). *Hero Twins* [digital art, Procreate]. College Board. *2023 AP Art and Design Exhibit.* https://apartanddesign.collegeboard.org/2023-student01

Hands, C., & Limniou, M. (2023). A longitudinal examination of student approaches to learning and metacognition. *Journal of Higher Education Theory and Practice, 23*(19). 125–150.

Holmes, N. (2017). Engaging with assessment: Increasing student engagement through continuous assessment. *Active Learning in Higher Education*. 19(1). 23–34.

Jagesic, S., Ewing, M., Feng, J., & Wyatt, J. (2020). *AP capstone™ participation, high school learning, and college outcomes: Early evidence* (Research Report No. RR 2020-09; ED603711). Educational Testing Service. https://eric.ed.gov/?id=ED603711

Jones, M. (2023). *Teacher Statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student06

Kuffel, D. (2023). *Teacher Statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student03

Lemp, E. (2023). *Teacher Statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student10

Nordfelt, A. (2023, December 1). *Currents* [Cone 10 clay, high fire glazes layered for custom effect, K9 & Las Vegas Red]. College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student05

Qualls, A. L. (1998). Culturally responsive assessment: Development strategies and validity issues. *The Journal of Negro Education*, *67*(3), 296–301. https://doi.org/10.2307/2668197

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28.* 4–13.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science. 18.* 119–144.

Stevens, G. (2023, December 1). *Teacher Statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student01

Stordahl, D. (2025, February 1). Caesar departs from Rome. *2024 AP Art and Design Exhibit.* https://apartanddesign.collegeboard.org/2024-student07

Taouki, I., Lallier, M., & Soto, D. (2022). The role of metacognition in monitoring performance and regulating learning in early readers. *Metacognition and Learning. 17.* 921–948.

Tian, E. (2023, December 1). *Flooding* [Watercolor and Pen on cut-up watercolor paper, collaged with glue.]. College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student14

Wiliam, D. (2011). *Formative assessment: Definitions and relationships.* Institute of Education, University of London. 1–26.

Wyatt, J., Feng, J., & Ewing, M. (2020, December). *AP® computer science principles and the STEM and computer science pipelines* [PDF]. AP Central. https://apcentral.collegeboard.org/media/pdf/ap-csp-and-stem-cs-pipelines.pdf

Zimmerman, S. *Teacher Statement.* College Board. 2023 AP Art and Design Exhibit. https://apartanddesign.collegeboard.org/2023-student11

Appendix A:
AP 2-D/3-D/Drawing Art and Design: 2024 Scoring Guidelines

Sustained Investigation Rubric

General Scoring Note

When applying the rubric, the score for each row should be considered independently from the other rows. You should award the score for that row based solely upon the criteria indicated, according to the **preponderance of evidence.** Student work may receive different scores for each row.

Each row includes decision rules and scoring notes used during the AP Art and Design Reading. Begin with score point 1 when applying the decision rules.

Row	Scoring Criteria		
A	**Inquiry** **Writing Prompt 1:** Identify the inquiry that guided your sustained investigation.		
	1	**2**	**3**
	Written evidence **does not identify** an inquiry.	***Written*** evidence **identifies** an inquiry **AND** ***Visual*** evidence **demonstrates** the inquiry.	***Written*** evidence **identifies** an inquiry. **AND** ***Visual*** evidence **demonstrates** the inquiry. **AND** The inquiry **guides** the development of the sustained investigation.
	Decision Rules and Scoring Notes *Read the student response to **writing prompt 1**.*		
	Does the ***written*** evidence **identify** an inquiry by describing discovery and exploration? (A question or a statement that merely identifies a theme or a topic is not an inquiry.) If no, award 1 point. If yes, move to criteria for score point 2.	Does the ***visual*** evidence **demonstrate** the inquiry? If no, award 1 point. If yes, move to criteria for score point 3.	Does the inquiry **guide** the development of the sustained investigation? If no, award 2 points. If yes, award 3 points.

B	Practice, Experimentation, and Revision **Writing Prompt 2:** Describe ways your sustained investigation developed through practice, experimentation, and revision.		
	1	**2**	**3**
	Visual evidence of practice, experimentation, and revision **does not relate** to a sustained investigation.	***Visual*** and ***written*** evidence of practice, experimentation, and revision **relates** to a sustained investigation.	***Visual*** evidence of practice, experimentation, and revision **demonstrates development** of the sustained investigation. **AND** ***Written*** evidence **describes ways** the sustained investigation **developed** through practice, experimentation, and revision.
	Decision Rules and Scoring Notes *Read the student response to* ***writing prompt 2.***		
	Is there ***visual*** evidence of practice, experimentation, and revision? **AND** Does the ***visual*** evidence of practice, experimentation, and revision **relate** to a sustained investigation? If no (for either or both), award 1 point. If yes (for both), move to criteria for score point 2.	Does the ***written*** evidence of practice, experimentation, and revision **relate** to a sustained investigation? If no, award 1 point. If yes, move to criteria for score point 3.	Does the *visual* evidence of practice, experimentation, and revision **demonstrate development** of the sustained investigation? **AND** Does the ***written*** evidence **describe ways** the sustained investigation **developed** through practice, experimentation, and revision? If no (for either or both), award 2 points. If yes (for both), award 3 points.

C	Materials, Processes, and Ideas		
	1	2	3
	Little to no evidence of **visual relationships** among materials, processes, and ideas.	**Visual relationships** among materials, processes, and ideas are **evident**.	**Visual relationships** among materials, processes, and ideas are **evident** and **demonstrate synthesis**.
	Decision Rules and Scoring Notes *In this row, written evidence is not scored but reading student responses may inform the evidence of visual relationships.*		
	Is there evidence of **visual relationships** among materials, processes, and ideas? If no, award 1 point If yes, move to criteria for score point 2.	Do the **visual relationships** among materials, processes, and ideas **demonstrate synthesis**? If no, award 2 points. If yes, award 3 points.	

D	2-D/3-D/ Drawing Skills		
	1	2	3
	Visual evidence of **rudimentary and moderate** 2-D/3-D/Drawing skills.	***Visual*** evidence of **moderate and good** 2-D/3-D/Drawing skills.	***Visual*** evidence of **good and advanced** 2-D/3-D/Drawing skills.
	Decision Rules and Scoring Notes		
	Does the ***visual*** evidence include **some** works with **good** (proficient) skills? If no, award 1 point. If yes, move to criteria for score point 2.	Does the ***visual*** evidence include **some** works with **advanced** (highly developed) skills? If no, award 2 points. If yes, move to criteria for score point 3.	Does the ***visual*** evidence across **all** works include **a range of good** to **advanced** skills? If no, award 2 points. If yes, award 3 points.

AP Art and Design Sustained Investigation Rubric Terminology (in alphabetical order)
2-D Art and Design Skills: The application of two-dimensional elements and principles—point, line, shape, plane, layer, form, space, texture, color, value, opacity, transparency, time; unity, variety, rhythm, movement, proportion, scale, balance, emphasis, contrast, repetition, figure/ground relationship, connection, juxtaposition, hierarchy
3-D Art and Design Skills: The application of three-dimensional elements and principles—point, line, shape, plane, layer, form, volume, mass, occupied/unoccupied space, texture, color, value, opacity, transparency, time; unity, variety, rhythm, movement, proportion, scale, balance, emphasis, contrast, repetition, connection, juxtaposition, hierarchy
Advanced: Highly developed
Demonstrate: To clearly show
Describe: Using words to communicate relevant information
Development: The furthering or advancing of an inquiry in a sustained investigation (through in-depth exploration of materials, processes, and ideas)
Discovery: To learn something through the process of making
Drawing Skills: The application of mark-making, line, surface, space, light and shade, composition
Experimentation: testing materials, processes, and/or ideas
Exploration: A journey of experimentation and discovery directed by inquiry
Evidence: To make obvious, seen, or understood
Good: proficient
Guides: The inquiry leads the process of making works of art and design
Ideas: Concepts used to make works of art and design (evident visually or in writing)

Identify: Indicate or provide information
Inquiry: The intentional process of questioning to guide exploration and discovery over time
Intent: The purpose or reason for exploring an idea
Materials: Physical substances used to make works of art and design
Moderate: Adequate
Practice: The repeated use of materials, processes, and/or ideas
Processes: Physical and conceptual activities including applications involved with making works of art and design
Questioning: Purposeful investigation and discovery in relationship to an idea
Reimagine: Reinterpret with imagination; rethink
Relate: Having a relationship and/or connection between
Revision: To modify, clarify, or reimagine works and ideas
Rudimentary: Emerging or undeveloped
Sustained Investigation: An inquiry-based and in-depth study of materials, processes, and ideas over time
Synthesis: Coalescence/integration of materials, processes, and ideas
Visual Evidence: The visual components that make up the student's works of art and design
Visual Relationships: Connections between the visual components included in a student's works of art and design
Ways: A series of actions or events leading in a direction or toward an objective
Written Evidence: The written components that accompany the student's works of art and design

Appendix B:

AP 2-D/3-D/Drawing Art and Design: 2024 Scoring Guidelines

Selected Works Rubric

General Scoring Note

When applying the rubric, you should award the score according to the **preponderance of evidence**; the response may not meet all three criteria indicated. However, if the written evidence is completely unrelated to the works, the **maximum** possible score is 2.

Scoring Criteria				
A. Written Evidence **B. 2-D/3-D/Drawing Skills** **C. Materials, Processes, and Ideas**				
The Selected Works demonstrate				
1	**2**	**3**	**4**	**5**
A. Written evidence **may identify** materials, processes, and ideas. B. **Little to no** visual evidence of 2-D/3-D/ Drawing skills. C. **Little to no** evidence of **visual relationships** among materials, processes, and ideas.	A. Written evidence **may identify** materials, processes, and ideas**.** B. Visual evidence of **rudimentary** 2-D/3-D/ Drawing skills. C. **Little to no** evidence of **visual relationships** among materials, processes, and ideas	A. Written evidence **identifies** materials, processes, and ideas. B. Visual evidence of **moderate** 2-D/3-D/Drawing skills. C. **Visual relationships** among materials, processes, and ideas are **evident** but **may be unclear** or **inconsistently demonstrated.**	A. Written evidence **identifies** materials, processes, and ideas. B. Visual evidence of **good** 2-D/3-D/ Drawing skills. C. **Visual relationships** among materials, processes, and ideas are **evident.**	A. Written evidence **identifies** materials, processes, and ideas. B. Visual evidence of **advanced** 2-D/3-D/ Drawing skills. C. **Visual relationships** among materials, processes, and ideas are **evident** and **demonstrate synthesis.**

Decision Rules and Scoring Notes				
A. Review written evidence:				
If the written evidence does not identify materials, processes, and ideas, the portfolio is only eligible for score points 1 and 2. If the written evidence identifies materials, processes, and ideas, the portfolio is eligible for all five score points.				
B. Review the application of 2-D/3-D/Drawing art and design skills to determine accomplishment level:				
1 Not present or unclear	**2** Emerging and undeveloped	**3** Adequate	**4** Proficient	**5** Highly Developed
C. Read the written evidence and then evaluate the visual relationships among materials, processes, and ideas:				
1 Little to none	**2** Little to none	**3** Evident, but unclear or inconsistently demonstrated	**4** Evident	**5** Evident and demonstrates synthesis

AP Art and Design Selected Works Rubric Terminology (in alphabetical order)
2-D Art and Design Skills: the application of two-dimensional elements and principles—point, line, shape, plane, layer, form, space, texture, color, value, opacity, transparency, time; unity, variety, rhythm, movement, proportion, scale, balance, emphasis, contrast, repetition, figure/ground relationship, connection, juxtaposition, hierarchy
3-D Art and Design Skills: the application of three-dimensional elements and principles—point, line, shape, plane, layer, form, volume, mass, occupied/unoccupied space, texture, color, value, opacity, transparency, time; unity, variety, rhythm, movement, proportion, scale, balance, emphasis, contrast, repetition, connection, juxtaposition, hierarchy
Advanced: highly developed
Demonstrate: to clearly show
Drawing Skills: the application of mark-making, line, surface, space, light and shade, composition
Evidence: to make obvious, seen, or understood
Good: proficient
Ideas: concepts used to make works of art and design (evident visually or in writing)
Identify: indicate or provide information
Inconsistent: not demonstrated in the same way or to the same degree across works of art and design
Materials: physical substances used to make works of art and design
Moderate: adequate
Processes: physical and conceptual activities involved with making works of art and design
Rudimentary: emerging or undeveloped
Selected Works: works of art that demonstrate synthesis of materials, processes, and ideas using 2-D/3-D/Drawing skills
Synthesis: coalescence/integration of materials, processes, and ideas
Unclear: not easily observable, discernable, or legible
Visual Evidence: the visual components that make up the student's works of art and design
Visual Relationships: connections between the visual components included in a student's works of art and design
Written Evidence: the written components that accompany the student's works of art and design

Research & Development Contributions to Assessment, Learning, Games, and Technology

Eva L. Baker and Gregory K. W. K. Chung

Abstract

This chapter presents a survey of illustrative examples of CRESST's R&D contributions to assessment, learning, games, and technology. The mission of CRESST was to understand the meaning of educational quality, including approaches involving evaluation and assessment. Examples from four major areas of R&D are presented: studies of writing assessment, the assessment of rifle marksmanship, evaluation of artificial intelligence systems, and game-based learning and assessment. A foundational element of the R&D was the exploration of assessment design, development, and validation in the context of learning, both as supporting the attainment of learning goals and as an outcome measure. Every example includes the importance of designing assessments to map to the purpose of evaluation and to provide as much transparency as possible. The examples illustrate the Handbook principles of *transparency*, *purpose* and *focus*, and *validity*.

Author Note

Eva L. Baker, ORCID: https://orcid.org/0000-0001-7347-2170

Gregory K. W. K. Chung, ORCID: https://orcid.org/0000-0003-4380-5661

We have no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Eva Baker, 300 Charles E. Young Drive North, SE&IS Building, Room 300, Box 951522, Los Angeles, CA 90095–1522.
Email: eva@ucla.edu

There was a time within memory when educational research and development was embraced as both important to develop new knowledge in the education and training world and for use as a scientific resource for the development of new applications intended to solve persistent problems. This chapter will highlight a few of the many contributions of the community, but it is tightly limited to a selection of work conducted at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). We describe four examples of programmatic research that took place over multiple years supported by the U.S. Departments of Education and Defense augmented by private support. The examples demonstrate CRESST's long-term commitment to designing assessments that uphold the core Assessment in the Service of Learning (AISL) principles of *transparency*, *purpose* and *focus*, and *validity*. The examples will also illustrate that developing assessment in the service of learning is not a new or abstract ideal for CRESST, but a throughline that has guided its work for decades.

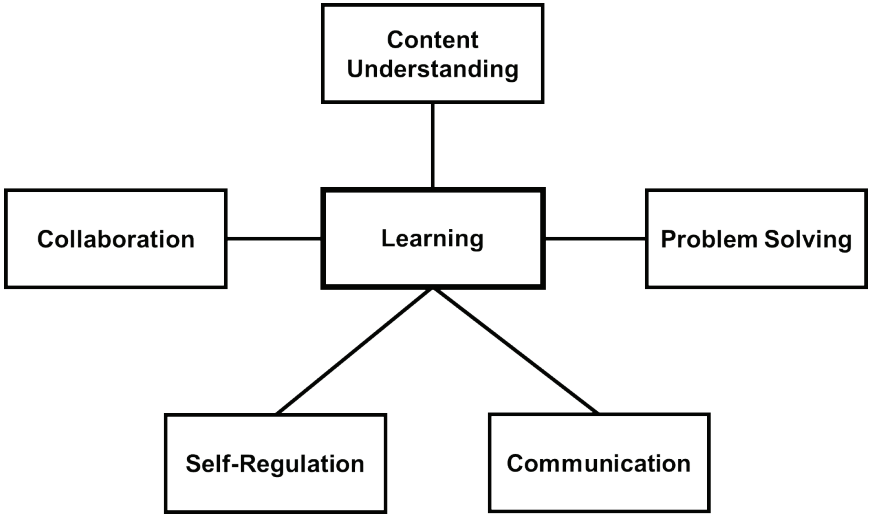
CRESST was originally developed in the mid-1960s as the Office of Education (prior to the inception of the United States Department of Education) responded to the reauthorization of Title I of the Elementary and Secondary Education Act. The response was a competition for a network of topically focused Research and Development Centers and a Network of Regional Education Laboratories focused on translation and development of usable educational options. UCLA received the 5-year award to focus on evaluation and supporting measurement and methodology in 1966 as the Center for the Study of Evaluation (CSE). Because these awards were developed to optimize the creativity of the scholars in the field, there was considerable latitude given to the design and management of research and development. When the Center grants were recompeted in 1984, CRESST was formally funded as a composite Center, where the focus was on assessment for use in schools, and partners of UCLA included universities, such as the Universities of Colorado, Illinois, and Stanford. CRESST also augmented its award with resources from state, local, federal, and private organizations.

Context of CRESST Research Design

During the period in which the research programs in this chapter occurred, three important conditions prevailed. First, the management of CRESST had extensive flexibility to select, compete, and conduct its research along with its scholars and students. It also was able to modify and adapt its objectives and procedures with little interference from the funding agencies. The ability to follow the directions of findings and to revise ongoing research plans is almost unheard of within recent funding from the federal government and as it may be in the future. Second, CRESST was a mission-focused organization. The mission of CRESST was to understand the meaning of educational quality including approaches involving evaluation and assessment. Technical studies to improve the scientific and statistical basis of the mission were an important concern, as was the exploration of alternatives to prevailing assessment approaches for policy uses. The third important element was to explore assessment in the context of learning, both as it supported the attainment of goals and as an outcome measure. In these efforts we collaborated with state and local agencies and specific organizations in the Department of Defense, including training for Army, Navy, and Marine Corps personnel.

A general model for the development of assessments was proposed and evolved over the years (Baker, 2007). Its central focus was learning supported by the various cognitive and domain requirements to promote the growth of learners. The original model, from Baker (1997), is displayed in Figure 1.

Figure 1.
Areas of Learning Identified for Model-Based Assessment (Baker, 1997)



The notion of the model also derived from research in computer science. This model was meant to be of general purpose and to be implemented in a variety of subject-matter domains. The idea of a general implementation, rather than an assessment approach that started with the subject matter, was a point of departure from traditional practice. Over the years, CRESST continued to develop and elaborate the model, for instance, using ontologies (Baker, 2007, 2012) to set boundaries for both subject matters to be included as well as the forms in which problem-solving would occur. Three criteria were developed to evaluate the quality of assessment: validity, utility, and credibility, all operating within an expectation of fairness and transparency.

The Examples

We include four examples of assessment and evaluation projects that had long-term programmatic reach. In each, we underscore the importance of learning and an understanding of both expert and learner perspectives. The principles animating this Handbook are also in play and include *transparency, purpose*

and *focus*, and *validity*. The first example we present is an effort that began with history assessment and developed into a writing assessment approach that was of general use. The second is the development of an approach to measure rifle marksmanship knowledge and skills. Both areas used expert performance as a criterion of quality as well as created models of transparent infrastructure that could be used in other assessment requirements. They were intended to focus simultaneously on learning-based assessments and outcome performance in a transparent manner. The third project focused on the development and evaluation of early versions of artificial intelligence including expert systems, natural language, and vision implementations using human benchmarking to measure the progress of AI systems. Our work also evaluated intelligent tutoring, games, and simulations. The fourth area extended R&D in learning game development and evaluation.

Simultaneously, CRESST was engaged in work in policy domains connected to local, state, federal, and international organizations focused on improving assessments, and their clarity, connection to learning and instruction, and attainment of learning goals.

Studies on Writing Assessment

This section will describe the R&D undertaken by CRESST in writing assessment. Its purpose was to apply our assessment model and develop a usable framework for the design and implementation of writing tasks to be used both in instruction and assessment of outcomes, and ultimately was generalized to other forms of constructed responses. The work involved emphases on the development of tasks to support the knowledge needed by students for writing and the ways in which scoring rubrics could be transparently designed to describe and to foster learning to write through feedback. CRESST began its interest in writing assessment in the late 1970s and focused on designs to assist state assessment agencies and to support an international study of written composition (Gorman et al., 1988). Around that time there were efforts by the Bay Area Writing Project (bawp.berkeley.edu), later the National Writing Project (www.nwp.org), to modify the way in which writing instruction took place, that is, to emphasize the process of planning, drafting, and revision. This approach also ultimately became an important part of classroom practice and assessment.

Writing Task Design: Prompt Development Supporting Prior Knowledge

We believed the writing process was only part of the solution, for our analyses and experience suggested that the design of writing tasks was not at all transparent or focused on student background. For essays to be used to evaluate content understanding, an approach was needed to capture students' prior experience. From our earlier studies, we had become convinced that students could not write well about topics on which they had little prior knowledge and that writing was not principally about appropriate style, organization, and mechanics, like punctuation and grammar, but about communicating, an approach supported by the work of Scardamalia et al. (1984). At early meetings of the IEA study on Written Composition (Gorman et al., 1988), we learned that colleagues provided content resources to writers to equalize prior knowledge and to help them flesh out their writing. CRESST staff eventually helped design tasks and scoring systems for the IEA research (Baker, 1982; Baker & Quellmalz, 1986). When CRESST was tasked by the federal government to develop secondary school history assessments, we chose to use writing as the scalable response mode to measure domain understanding. Starting with 10th grade U.S. history, we began an analysis of that content included in popular textbooks to understand student knowledge to be assessed. Unfortunately, we discovered that the treatments of important topics, such as the causes of the Civil War, were presented superficially in a paragraph of text or two and could at best provide the learner with only a thin layer of knowledge. Modeling the IEA R&D, we provided the learners with relatively short primary sources from the period of interest, using contrasting positions of politicians, for instance, the debate speeches by Abraham Lincoln and Stephen Douglas. We followed this model using opposing letters or speeches for the Revolutionary period, the Civil War, immigration in the early 20th century, and World War II among other key events in U.S. history.

Students were to read the given primary sources and then to write an essay in letter form to an absent classmate explaining the meaning of the contrasting positions. Note that over the years, we created similar assessment tasks using primary sources in history, geography, social studies, multidisciplinary topics, and science, where students read about situations and experiments rather than contrasting positions (Baker et al., 1990). In one scaled effort, we applied this approach to statewide trials in the state of Hawaii, using content in Hawaiian history and social studies topics for younger students in upper elementary school (Baker et al., 1991, 1996).

Improving on Scoring Approaches

Simultaneously, the team embarked on approaches to improve scoring by making it more transparent and valid. As noted, our interests were both outcome measures and essays assigned during courses. In both cases, the task was to improve the quality and validity of the scoring, to focus on elements that could be used for student feedback, and to reduce the time burden on teachers that scoring assigned essays imposed. The last point was critical because we had learned that teachers often severely limited the number of writing assignments given to students simply because they had no time to evaluate them. We intended to find evaluation approaches that got to the core of performance without requiring the traditional annotation and lengthy comments by teachers. Moreover, there were also approaches at the time that argued that every writing assignment required its own scoring rubric (See for example, Graves, 1978). While the idea of extracting specific information for each assignment made some sense, the reality was that teachers having to learn to use a different scoring rubric for each assignment was an incredibly unlikely outcome. Idiosyncratic scoring regimes also inhibited the ability to monitor student growth in performance over time, where a common criterion is desirable.

Do What I Do, Not What I Say

At CRESST, we decided to explore how the design of scoring rubrics could move beyond teachers' agreed-upon preferences. Our question was simple: Could we make inferences from the actual writing of experts to determine criteria for scoring student work? To that end, we asked teachers and other history experts in graduate school to write answers to prompts about epochs in U.S. history using the provided contrasting speeches. Careful analysis of the experts' writing found they organized their answers using principles or themes, they brought to bear prior knowledge external to that in the provided prompts, they used concrete examples to support their position often from the provided resources, and they avoided major mistakes or misconceptions. To use models of expertise proposed by renowned cognitive researchers (e.g., Chi et al., 1988; Ericsson & Charness, 1994; Gentner & Genter, 1983) we conducted expert-novice studies to confirm common elements in expert writing. An additional set of research involved developing and validating rater training (Quellmalz, 1982) where we focused on accuracy and speed, as we wished to support opportunities for more writing for students.

Impact and Future

The consequences of our work resulted in the development of writing approaches used for a number of state assessments, NAEP (Baker, 1981; Baker et al., 1986), and for multiyear work across literacy and mathematics domains at the elementary school level in the Los Angeles Unified School District (Niemi & Baker, 1998). We also applied these analyses to the evaluation of A level writing in Great Britain (Baker et al., 2002). Current work in AI scoring should include models generated by expert raters rather than simply interpreting identified rubrics. Our current work has focused on the identification of assessment tasks using AI-defined ontologies and domain task generation.

One of the most enduring outcomes of the studies on writing was the generality and utility of the CRESST assessment model and its emphasis on starting with learners and learning outcomes to drive the design of assessments and measures at CRESST (Baker, 1997, 2007; Baker & Gordon, 2014; Baker et al., 2022; O'Neil et al., 1990).

Assessment of Rifle Marksmanship

One of the most remarkable achievements in United States Marine Corps (USMC) marksmanship training is in developing a shooter's skill to routinely hit a 19-inch circular area at 500 yards in the prone position. The challenge posed to CRESST was to develop a way to assess marksmanship *in a distance learning context* with the goal of helping the USMC improve their non-infantry Marines' marksmanship skills.

In order to develop assessments of what was commonly believed at the time essentially a motor task, without being able to directly observe the shooter carrying out the task, required CRESST to start a program of research from first principles. Many of the methodologies developed for writing assessment were adapted for marksmanship. New frameworks and technologies needed to be developed as well, as marksmanship was never studied from an assessment perspective. In the remainder of this example, we describe the R&D program and illustrate how the domain of marksmanship was defined, how the measures were developed and validated, and how novel measurement approaches were used to explore individualizing instruction.

Determinants of Marksmanship Reexamined

At the start of the research, the marksmanship literature was focused almost exclusively on the proper execution of the motor aspects of the factors needed to establish a stable platform for the rifle and the components that underlie aiming. There was almost no conceptualization of marksmanship as a complex skill and little research to draw on to form a coherent assessment framework. To develop assessments of marksmanship that could operate under distance learning conditions, we needed to understand the underlying factors external and internal to the shooter that affected marksmanship performance.

Based on the literature and interviews with subject-matter experts (SMEs), we decomposed marksmanship performance as a function of factors within the purview of the shooter (perceptual-motor, cognitive, affective) and external to the shooter (weather, equipment). This conceptualization mirrored the CRESST assessment model (Baker, 1997, 2007) (See Figure 1). While the individual components of the model differed, how the components were identified and the role of the components as the focus for the assessments remained the same.

A key contribution was incorporating cognitive and affective components into the research. By conceptualizing marksmanship as a complex skill, we could rely on a skill acquisition model to understand how knowledge and performance interacted over time (Ackerman, 1987, 1992; Fitts & Posner, 1967). Skill development is believed to move from a learning phase to a practice phase and then to an automaticity phase. When applied to marksmanship, trainees in the learning phase are attempting to learn the concepts and rules of marksmanship. Trainees in the practice phase know what to do and practice implementing the various rules and procedures. Trainees in the automaticity phase can smoothly execute the skill with little overt consideration of the rules and procedures.

The skill model predicted the poorest performance during the learning phase when trainees are least likely to have acquired and internalized the knowledge required to shoot well (i.e., Marines who do not routinely handle weapons), suggesting measures of knowledge might be the most sensitive. For trainees in the practice and automaticity phases, perceptual-motor measures could be expected to be stronger predictors of performance. Given our population was non-infantry entry- and sustainment-level Marines, we focused on developing assessments for trainees in the learning phase and with the constraint that the assessments would need to work in a distance learning context.

Assessment Development and Validation

While we had a theoretical model of how skill develops and which phase of skill development to focus on, we needed to know precisely what knowledge Marines needed to know, how this knowledge related to shooting performance, and whether this knowledge was malleable (i.e., for applications in future distance learning training applications).

We used the CRESST assessment model to guide assessment development. We focus on identifying the cognitive demands that bear on learning, and these cognitive demands drive the design of the assessment task. The model led us to ask three questions: What are the processes (cognitive, affective, motor) that influence a trainee's successful execution of a task? What are the most direct ways of observing and measuring those processes without the measures altering the measurement itself? and How can these measures be validated to support the inferences drawn from the scores?

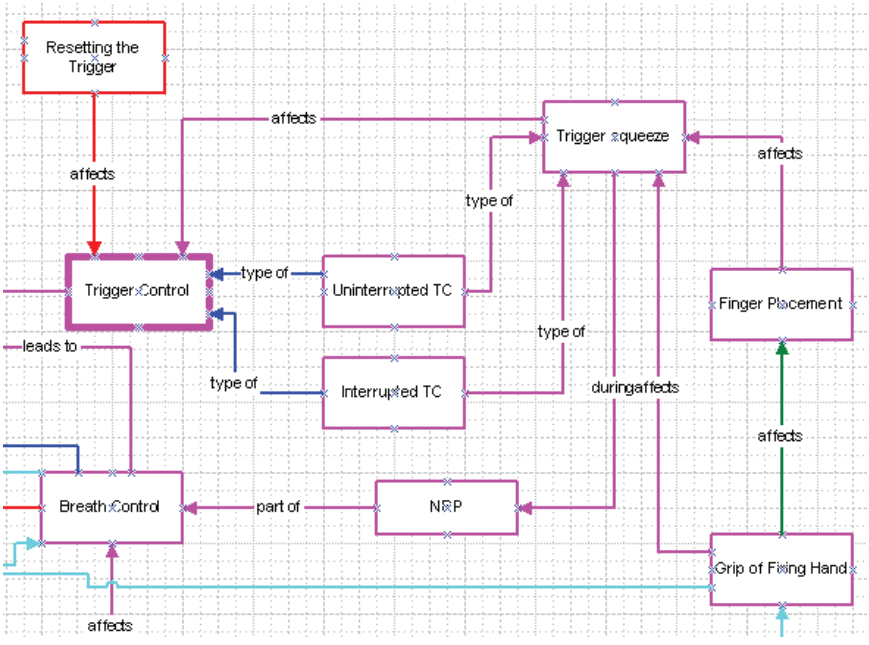
Knowledge Representations

We relied extensively on knowledge representations for practical reasons. Knowledge mapping, a method developed in the writing assessment studies to measure conceptual knowledge (Herl et al., 1996), was used to capture experts' understanding of the domain (Chung, Michiuye, et al., 2002). Experts tend to represent only the most important ideas in a domain, which is an efficient way to identify the major topic areas for an assessment. We also culled from field manuals specific cause-effect relations to augment experts' knowledge maps. The knowledge elements from experts and field manuals were stored in an ontology that was later used for scoring purposes and for instructional purposes.

Capturing Experts' Knowledge

USMC coaches and a scout sniper served as SMEs. Each SME created a knowledge map to represent how they viewed the relations among the various concepts. Figure 2 shows a fragment of the knowledge map. When we overlaid the different experts' maps, it was clear that the most sophisticated map was from a scout sniper. His map spanned multiple areas of marksmanship, reflected what we were learning from SME interviews, and presented an integrated theory of marksmanship. The differences among the various maps were consistent with USMC training, where scout snipers, compared to coaches, receive far more comprehensive and in-depth training on marksmanship.

Figure 2.
Fragment of Experts' Knowledge Maps of Rifle Marksmanship



Measures of Rifle Marksmanship Knowledge

The combination of USMC field manuals (e.g., USMC, 2001), expert interviews, and follow-up discussions with the SMEs made it clear that there was a strong knowledge component to marksmanship in addition to perceptual-motor skills. We organized this knowledge into a framework for rifle marksmanship composed of the following components: cognitive (e.g., domain knowledge), perceptual-motor (aiming, sight picture, fine and gross motor), affective (e.g., anxiety), and equipment and weather.

The set of measures we developed addressed the different components of rifle marksmanship: (a) a broad measure of marksmanship knowledge that sampled the domain and used a selected-response format; (b) a measure of conceptual knowledge using knowledge mapping; (c) an interactive task asking shooters to

identify proper and improper position elements; (d) an interactive task to interpret shot group patterns; and (e) questionnaires to survey trainees' worry, anxiety, and firing line experience. The measures went through multiple reviews by our SMEs.

Validation of Rifle Marksmanship Measures

Empirical validation tested the measures on samples with different levels of experience (non-infantry entry- or sustainment-level Marines and marksmanship coaches; high and low shooting performance) and aptitude (officer candidate school), and on trainees prior to and after instruction. In a series of three studies, we gathered evidence that, in general, suggested that the knowledge measures were sensitive to instruction, predicted record-fire scores moderately in less experienced samples, and when combined with other variables within the skill acquisition framework, predicted record-fire scores as well as scores from a rifle simulator (Chung et al., 2004). We next briefly discuss two interesting measures used in the marksmanship research: knowledge mapping and self-reported worry and anxiety.

While it was clear from the writing assessment studies that knowledge maps could be used to assess conceptual knowledge, knowledge maps were never used in a military training context. As in Herl et al. (1996), experts' maps were used as criterion maps against which trainee maps were scored. We found knowledge maps were sensitive to instruction and sensitive to expertise. Marines' knowledge map scores increased over the course of instruction (Chung et al., 2004, Study 2, 3) and Marines with more marksmanship experience scored higher than those with less experience (Chung et al., 2004, Study 2). These results are consistent with other studies that tested knowledge maps for instructional sensitivity and expert-novice differences (e.g., Herl et al., 1996, 1999; Ruiz-Primo et al., 2001).

The role of anxiety on marksmanship performance was recognized over 100 years ago. Gates (1918) reported that novice shooters' performance was affected severely by their dwelling on steadiness factors (e.g., uttering "There, I moved again"; p. 3). In our studies, the state measures of worry and anxiety administered on qualification day were among the highest predictors of record-fire score, with state anxiety and worry significantly and negatively correlating with record-fire scores (*r*s ranging from -.4 to -.5) (Chung et al., 2004, Study 2; 2005). Furthermore, when we tested the joint effects of aptitude and state worry inspired by Ackerman's (1987, 1992) study of how aptitude influences performance during the learning

phase, we found that aptitude and state worry predicted record-fire scores with a multiple *R* of .67, with state worry accounting for 34% of the variance and aptitude accounting for 11% (Chung et al., 2005).

Using Assessment to Improve Learning

Because one of our requirements was to develop assessments for a distance learning context, we anticipated the need to demonstrate how assessment information could be used for training purposes. Thus, we developed several methodologies to support future distance learning training applications given the widespread interest in the military in individualizing instruction (Bewley et al., 2009). One of the most important methodologies was the use of knowledge representations or ontologies. An ontology is domain knowledge expressed as a set of concepts and the relations that hold among the concepts (Baker, 2012; Chung et al., 2003; Gruber, 1995). Because ontologies are machine-readable and structured, software can be developed to operate on them. In our case, we created an ontology to represent marksmanship knowledge and linked instructional content in the form of text, figures, and video snippets from USMC training videos to a marksmanship concept (Chung et al., 2004). We then tested on a small sample whether individualizing instruction was effective. The results suggested that Marines receiving individualized instruction improved on topics where they initially had a knowledge gap and not on concepts they did not receive instruction on. The study strongly suggested that the methods used to model knowledge, assess knowledge, and tailor instruction were promising (Chung et al., 2003).

While we could measure one's knowledge of how to carry out a procedure (e.g., trigger control), we had no way to directly measure the execution of that skill. Our follow-on marksmanship R&D work, funded by the Defense Advanced Research Projects Agency (DARPA) investigated whether we could accelerate the acquisition of marksmanship skills. We used sensors to gather information on the difficult-to-observe processes of breath control, trigger control, and muzzle wobble (Espinosa et al., 2009; Nagashima et al., 2009) and we used an observation checklist of the various position elements considered important by experts and USMC doctrine. We tested whether we could use these fine-grained measures to (a) diagnose the novice participants' shooting problems and (b) provide effective individualized remediation using brief video-based instruction. We modeled experts' shots using the sensor data and were able to classify each

shot as expert-like or not (Nagashima et al., 2009). We found that participants who received tailored remediation significantly outperformed those who did not receive tailored instruction, with an average of 2.0 (out of 5) expert-like shots (vs. 1.0 expert-like shots). While this result may seem minor, improving novices' ability to better execute a complex skill composed of cognitive, affective, and perceptual-motor factors in 65 minutes suggested a potentially efficient approach (Chung et al., 2008).

Impact

The idea that rifle marksmanship comprises cognitive, affective, and perceptual-motor factors was novel at the time. The notion that marksmanship has a cognitive component and is a complex skill appears to be accepted by researchers worldwide as evidenced by citations to our work. The insight that marksmanship had a cognitive component was a natural development given CRESST's approach to assessment design best exemplified by Baker's (1974, 1997, 2007) focus on cognition and validity. By grounding the measurement effort around cognition and skill development, new insights were gained about which kinds of assessments would be appropriate for trainees depending on their skill development. This tailoring of measures and content was carried into instructional applications in math (e.g., Chung, Delacruz, et al., 2016), further demonstrating the utility and generality of focusing on cognitive demands first and foremost.

The second impact was the tools and methods developed or applied during the course of the research. Capturing SMEs' knowledge representation served as a method to distill the most important ideas of a domain and a way to assess learners' conceptual knowledge. The use of hardware sensors for measurement purposes would continue (e.g., Chung et al., 2021), and the conceptual and practical connection between measurement and instruction would continue to influence CRESST's technology-based R&D.

Evaluation of Artificial Intelligence (AI) Systems

AI is now at the center of attention in learning technology. We will describe a series of encounters with AI-based systems, for the most part seeking to evaluate their effectiveness. Many studies resulted in a lack of definitive findings because of the limited power of early interventions. Nonetheless, early in CRESST's history, we began numerous studies of advanced technologies, using relatively primitive implementations to explore and evaluate consequences (Baker, 1988). The story of our evaluations of artificial intelligence (AI) systems includes a few pieces. A significant note is that our work was ahead of its time; that is, it stood apart from the usual technology studies in its oddness. Only now, as AI has penetrated the daily lives of many users, our ancient studies are of renewed interest. Our evaluations included early games and simulations, expert systems and models used to support natural language processing and vision systems, and intelligent tutoring systems to promote learning. An important side effect which we will describe is our use of aspects of intelligent system design to enhance our design and implementation of assessments.

AI Games, Simulations, and Intelligent Tutoring Systems

The first game we evaluated using AI was WEST, derived from How the West Was Won, and created by Richard Burton and John Seeley Brown (Burton & Brown, 1979), titans in the early development of AI. Fascinated by the early efforts in this area, CRESST obtained support from NASA to conduct the evaluation of the game along with the Jet Propulsion Laboratory. The principal AI option in the game was a coach which was to support students' learning. We dismantled the coach, and our experiment included students who were exposed to the game with and without the coach support. The findings did not support the utility of the coach.

A second effort was supported by DARPA and was two-pronged. One set of activities was to evaluate AI-based approaches to support former service members who were afflicted with post-traumatic stress disorder (PTSD). A few private companies had created options that could be accessed through smartphones and from periods of activity and other everyday behaviors could infer episodes of PTSD and then implement support. The difficulty with this approach was that it required long periods of use as well as permissions by the users for analyses of their daily technology use. The evaluation design and beginning implementation were carried

out, but the project eventually drew no conclusions because of few users who participated for the desired length of commitment (Baker et al., 2015).

The DARPA game study ENGAGE involved the evaluation of a game developed at Carnegie Mellon University. The game was developed for primary-school-aged learners and taught children to use an adaptation of balance scales to reach conclusions about equivalence (Aleven et al., 2013). Our major evaluation finding was that games could increase the self-efficacy of young learners in the topical subject matter (Baker, 2015; Baker et al., 2016).

As part of this work, CRESST developed its own game focused on physics for 6-year-olds. The game taught concepts of mass, acceleration, and friction, where students needed to manipulate the variables to allow a train to exactly reach its station. In addition, students were to deal with bullying that occurred among characters in the game. Again, limitations of the obtained data interfered with our inferences of effectiveness. We were able to implement and further develop a framework for the evaluation of games that included cognitive demands, domain knowledge, and detailed specifications (Baker et al., 2011; Baker & Delacruz, 2016). Moreover, in developing the scenarios for the physics game, we evolved an assessment design strategy useful for creating exchangeable performance assessments efficiently. The approach created "slots" for key variables in content, task, cognitive demand, and situation that allowed the generation of comparable tasks quickly and at low cost (Baker & Delacruz, 2008).

Simulations

One outcome of our R&D around the evaluation of simulations was the development of novel measures and approaches. Simulations provide learners with experiences that might not be feasible in a classroom or training setting. The simulations CRESST evaluated required learners to engage in problem-solving and reasoning, which also meant the need for measures that would be sensitive to these higher level learning outcomes.

A persistent design goal was to measure the phenomenon in as direct a way as possible. This objective pushed R&D developments in three areas: first, to continue to apply the CRESST model of assessment, which maintained our attention on how cognitive demands of the simulation task related to the assessment task design; second, to adopt or develop measures that reflected the productive (or

nonproductive) uses of the unique learning affordances of the simulations; and third, to instrument our evaluation tools to capture and log fine-grained learner-system interactions (also called log data, trace data, or clickstream) and to use those data for assessment purposes.

Evaluating Content Understanding and Problem-Solving

Beginning in the mid-1990s, we began to explore how simulations could be used for assessment purposes. We became increasingly confident over several studies that simulations that required performance demonstrations could also be used for assessment purposes. For example, we developed a simulated web environment to evaluate middle-school students' content understanding and problem-solving. Content understanding was measured with knowledge maps, and problem-solving was measured by information seeking and search (Baker & Mayer, 1999). The educational setting was the Department of Defense Education Activity (DoDEA) middle schools in Germany, where large investments in computer-aided educational tools were introduced into the schools. The study found students' search skills and knowledge of environmental science significantly improved from the fall to spring semesters and knowledge map scores were significantly related with the quality of their search behavior (*r*s from .4 to .5) (Schacter et al., 1999).

This study was foundational in that we demonstrated the technical feasibility of collecting fine-grained behavioral process data and showed that students' online behavior was related to their content understanding and problem-solving outcomes. The capability to link students' behavior to their improved knowledge led to an obvious understanding: If students attended to the relevant content, they would learn that content. While a simplistic insight and long known in the verbal memory research, this finding was with an educationally relevant task where we could directly tie learners' behavior to the to-be-learned content. The challenge was not in the technology development or instrumentation, but rather in being able to create tasks where the learner interaction was aligned with the cognitive demands that influenced outcome performance. We concluded that under this situation, behavioral process data could be highly informative.

Given the promising results of the web search study, we then examined another simulation to gather validity evidence of the degree to which learners' online behavior reflected their cognitive processes. This linkage was important to establish because there was scant evidence in the literature to confirm that

learners' online behaviors were representations of their thinking. Establishing such a link would increase our confidence in the use of online behavior as a source of evidence about learning processes. Chung et al. (2002) collected process data and concurrent think-alouds from students as they engaged in a web-based problem-solving simulation task. The simulation required learners to determine the parents of five children (Stevens et al., 1999). The learners could access information sources with different credibility (e.g., genetic lab test results, opinions of people, library) to rule out candidate parents.

Similar to the web search results (Schacter et al., 1999), task performance was significantly and positively related to learners' fine-grained behavior reflecting the use of credible sources and negatively related to use of non-credible sources. We also confirmed that productive cognitive processes (based on students' think-alouds) were significantly related to existing validated measures of reasoning. When we examined how learners' cognitive processes were related to their online behaviors, we found that productive cognitive processing was significantly associated with task performance and productive learner behaviors and vice versa, with the magnitude of correlations in the .5 to .7 range. The results of triangulating cognitive processes derived from think-alouds, validated measures of reasoning, and learners' behaviors bolstered considerably our confidence in the use of online behavioral data for measurement purposes (Chung, de Vries, et al., 2002).

The final simulation example addressed the extent to which a simulation designed specifically for training purposes could be used for assessment purposes (Iseli et al., 2019; Savitsky, 2013). For this study, CRESST developed and validated methods to assess both declarative and procedural skills for two ultrasound-guided procedures taught in the simulator. Declarative knowledge was measured by a general test of knowledge of the two ultrasound procedures. Procedural knowledge was measured by the quality of sonographers' ultrasound scanning with a probe. The probe-motion measures were derived from moment-to-moment telemetry of the pitch, yaw, and roll of the probe. We found that more experienced sonographers demonstrated superior overall task performance and probe manipulation skills compared to less experienced sonographers, with effect sizes between the two groups of participants ranging from 0.2 to 2.0 across the various probe-based measures. These results, coupled with the marksmanship study involving sensors, suggested that the data from hardware sensers could be used in similar ways as we were using online behavior data. These results also suggested

a kind of generality: The utility of learner behavioral data is less about the specific source (software or hardware) and much more about whether the behavior is a manifestation of cognitive processes of interest.

A major theme of our simulation evaluation examples is the use of the CRESST assessment model. In every study, the learner and learning outcomes were the focus of the assessment task design effort. The cognitive demands required of the task, and in particular the unique aspects of the simulation task, guided the development of novel assessments that measured as directly as possible the presumed learning outcomes and processes. The close attention to cognitive demands and how they manifest in learners in a given task design also led to insights about which kinds of behavior in the simulation carried information related to learning and which did not. These insights would be carried into future work on game-based learning and game-based measurement.

Intelligent Tutoring Systems (ITS)

One of the most common and early uses of AI was its application to intelligent systems for learning. Early called intelligent computer assisted instruction (ICAI), several studies were conducted by CRESST (O'Neil & Baker, 1987). About two decades later these inquiries continued, supported by the Office of Naval Research (Kumar et al., 2015; VanLehn et al., 2016). In this section, instead of presenting a full example of an ITS evaluation, we present an example of measures development, a key issue when evaluating systems that individualize instruction.

The results of any evaluation rest on the quality of the outcome and process measures. ITS presents a special case because the instruction tends to be individualized, and system instructional decisions are made using granular data (e.g., presenting feedback tailored to a specific type of learner response). Thus, a challenge posed by ITSs (and systems that individualize instruction) is determining effectiveness when different students receive different degrees of content exposure, practice, and feedback.

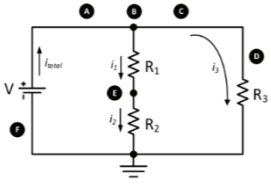
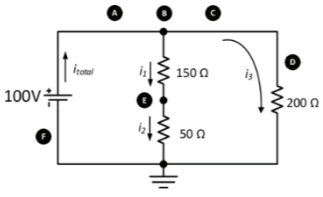
The approach we used focused on the precision of measurement. Because an ITS often attempts to remediate knowledge gaps on specific topics (e.g., understanding how to compute the equivalent resistance of three resistors in parallel), we reasoned that the measures used in evaluating the ITS should also match the precision of the instruction as a broader outcome measure might not

detect very narrow effects. One example of this approach was the evaluation of the ITS *LearnForm* (Kumar et al., 2015). *LearnForm* was an ITS problem-solving environment where students were first presented with a selected-response item. If they answered the item incorrectly, they could receive step-by-step, granular instruction and practice on the underlying topics related to the test item. The system's evaluation focused on electric circuits in AP Physics classrooms.

The measures development consisted of a physics SME first developing an ontology of electric circuits to identify the important domain concepts. These concepts were decomposed into specific knowledge components. Item development involved reviewing the electric circuit literature for misconceptions, developing canonical circuit topologies, and evaluating candidate items against the set of knowledge components.

Successful analysis of a circuit requires the simultaneous consideration of the relations among voltage, current, and resistance. To mirror this cognitive demand, we adapted an item format from Richardson et al. (1933, p. 55) and discussed in Haladyna and Rodriguez (2013). As shown in Figure 3, the item was used to assess conceptual understanding of the relations among current, voltage, and resistance, and procedural knowledge of how to apply Ohm's Law to compute voltage and current.

Figure 3.
Example Conceptual and Procedural Knowledge Items

Testing for conceptual knowledge	Testing for procedural knowledge
	
Example items	Example items [b]
1. The voltage at A is [>, =, <, indeterminate, don't know] [a] than the voltage at E.	1. The voltage at A = _____. 2. The voltage at E = _____.
2. i_1 is [>, =, <, indeterminate, don't know][a] than i_2.	3. i_1 = _____. 4. i_2 = _____.

[a] Participants select one option. [b] Participants compute the answer.

The format shown in Figure 3 allowed us to create seven scales with 41 items. The scales underwent multiple rounds of review and validation testing. The internal reliability of the scales (Cronbach's alphas) ranged from 0.7 to 0.8 (Chung, Madni, et al., 2014). Knowledge sensitivity was verified by comparing electrical engineering (EE) students to a general sample, where EE students performed significantly higher than the general sample. Instructional sensitivity of the scales was verified by first showing that the EE sample did not change over instruction (i.e., no difference in pretest and posttest scores), and also showing that scores increased from pretest to posttest in the general sample (*d*s ranging from 0.3 to 0.5). *LearnForm* effectiveness was demonstrated with an evaluation sample that improved from pretest to posttest on the scales (*d*s ranging from 0.7 to 1.9), and by demonstrating that learners who received the step-by-step instruction outperformed those who could opt out of the step-by-step instruction on the conceptual circuit analysis measure (*d* = 0.8) (Chung et al., 2015).

Human Benchmarking of AI Systems

DARPA supported an innovative set of studies evaluating early AI systems using human performance as the guide (Baker & Butler, 1991; Swigger et al., 1990). These systems included an example of natural language processing (NLP), a completed expert system in the area of scheduling, a vision system (Baker et al., 1988), and an expert systems shell. The project was initially and deliberately controversial in the computer science area, because the principal investigator was not a computer scientist. However, the evaluators of each major component came from the computer science domain. The question posed in this study was how well the system performed in comparison to human performance. Common tasks for humans were transformed and were acted upon by systems and then levels of performance were inferred. For instance, early evidence from NLP systems suggested at that time, performance was like that of a primary-school-aged child (Baker, 1994). For the most part, the work was conducted, albeit with interruptions from the funding agency when the initial supporter changed agencies. In the expert system scheduling analysis, systems managing scheduling of airplanes to gates existed, and similar tasks were given to people (O'Neil et al., 1994). Reports of this work were developed and form some of the basis of current studies of system predeveloped problem sets to evaluate comparatively the efficiency and growth of distributed systems such as ChatGPT (Baker, 1989; Baker et al., 2025).

Impact

To understand the implications of our early work in evaluating AI, two conditions are clear. One is that early formulations were extremely limited in design, and so were the evaluation options open to CRESST. To this day, CRESST is continuing to engage with AI options to support our own work in the design of ontologies and performance assessments for learning, to develop measures for various types of data collection, to explore the use of intelligent agents to act as simulated students for assessment and evaluation, and to attempt to understand what learning quality means in the era of expanding machine intelligence.

Game-Based Learning and Assessment

In this section, we present selected examples, findings, and insights from our R&D portfolio around games for learning and assessment. While the examples are drawn from our work sponsored by the U.S. Department of Education (ED), the Institute of Education Sciences (IES), and PBS KIDS, many of the methodologies and lessons learned were the result of continuous cross-fertilization among the various ongoing military games and simulation programs at CRESST sponsored by the Office of Naval Research (e.g., Baker & O'Neil, 2002; Iseli & Jha, 2016; Iseli et al., 2010; Koenig et al., 2010), DARPA (e.g., Baker et al., 2012; Baker & Delacruz, 2016; Madni et al., 2013; O'Neil et al., 2021), California Department of Education (e.g., Chung et al., 2018), private foundations (e.g., Chung, de Vries, et al., 2002), and start-up organizations (e.g., Ihlenfeldt et al., 2025).

Game-Based Learning

In 2009, CRESST was awarded a multimillion-dollar 5-year national R&D center on instructional technology grant from the U.S. Department of Education, Institute of Education Sciences (IES). The center, named the Center for Advanced Technology in Schools (CATS), developed and tested fractions math games for underperforming middle-school students in a cluster randomized controlled trial (RCT). The RCT involved 23 schools, 59 classrooms, and 1,468 students and demonstrated that students who played four fractions games performed higher on a test of fractions knowledge, compared to the comparison group who played four solving equations games (d = 0.23) (CATS, 2012; Chung et al., 2014; ED, IES, WWC, 2015). We next highlight several innovative aspects of CATS: coherent design process, game as testbed, gameplay as a data source, and advanced statistical modeling.

Coherent Design Process

We used the CRESST assessment model (Baker, 1997, 2007) to develop knowledge specifications. Ontologies were used to describe the major concepts and relations in the content domains (Baker, 2012) and the knowledge specifications succinctly described the target concepts, types of stimuli to elicit student responses, and performance expectations. The knowledge specifications standardized the requirements for assessment design, game design, and professional development for the target domains (rational number equivalence, CATS, 2013b; solving equations, CATS, 2013c; functions, CATS, 2013a). A fragment of the knowledge specification for rational number equivalence is shown in Figure 4.

Figure 4.
Snippet of Knowledge Specifications for Rational Number Equivalence

		Computational Fluency: Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures, and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
Rational Number Equivalence Knowledge Specifications		**When presented with… *(Assessment Stimulus)***	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
1.0.0. Does the student understand the importance of the unit whole or amount?					
	1.1.0. The size of a rational number is relative to how one Whole Unit is defined.	Any rational number	Place it on a number line relative to the whole interval explicitly (0 and 1 labeled) or implicitly (0 and an integer other than 1 labeled) defined.	Apparent contradictions involving rational number such as ¾ < ½ or ½ does not equal ½	Explain that the contradiction can be resolved if their relative wholes must be equal when comparing.
		A unit whole (interval, volume, area, etc.)	Show how much of the whole must be shaded to represent a fractional amount.		

All assessments, games and game levels, and professional development were designed against the knowledge specifications. Both the game levels and assessment items were mapped to the knowledge specifications, allowing verification of adequate domain coverage and alignment between the instruction, the game levels, and the assessment.

Game Testbed to Accelerate Research

A second innovation that enabled CRESST to conduct 17 design studies over two years was to design the games as a testbed. All games were designed to allow researchers to specify the level design using a text file instead of needing a programmer to program the levels. For example, in the game *Save Patch*, if a player failed the level, researchers could specify instruction or feedback tailored for the first failure, second failure, and so on, and also specify that the instruction be delivered in different modalities (e.g., text only, video). An example of the utility of the testbed was in simply modifying five text files to create five versions of *Save Patch* to identify the most promising forms of feedback to implement in the games used in the RCT (Vendlinski et al., 2011).

Gameplay as a Data Source

A third innovation was the use of fine-grained telemetry for measurement purposes. Our prior work with process data (Chung, de Vries, et al., 2002; Schacter et al., 1999) guided our telemetry design of what game mechanics to instrument, what game states to record, how to structure the data, and how to format and log the data. Yet we were unsure whether gameplay itself carried information about

learning as game-based learning was an emerging field at the time. While our first three experimental studies did not show outcome differences due to instructional variations, we did find significant gains over gameplay (*d*s from .3 to .4), hinting that the game design and game mechanics were effective in conveying the fractions concepts (Chung et al., 2010). We found that players receiving math-focused instruction (vs. game-focused instruction) generally committed fewer errors in the game that were related to math (*d*s from 0.3 to 0.5), and the math posttest was significantly related to gameplay behaviors reflecting successful fraction addition (*r*s from 0.3 to 0.6) and negatively related to gameplay behaviors reflecting unsuccessful fraction addition (*r*s around -.3). These results suggested that gameplay behavior itself carried information about learners' fractions knowledge.

These results were generally replicated in subsequent studies, suggesting that the game facilitated learners' acquisition of fractions knowledge (Vendlinski et al., 2011). Furthermore, the pattern of how gameplay related to tests of knowledge repeatedly showed that knowledge was positively related to productive gameplay behavior and negatively related to unproductive gameplay behavior, consistent with prior work (Chung & Baker, 2003; Chung, de Vries, et al., 2002; Schacter et al., 1999). These results spurred continued examination of the use of process data, including using data mining methods to detect misconceptions (e.g., Kerr, 2014; Kerr & Chung, 2012a, 2012b, 2013b), to test whether instructional variations affected specific gameplay behaviors (Buschang et al., 2012; Chung et al., 2010), to identify different learning trajectories (Kerr & Chung, 2013a), to model diagnostic assessments (Levy, 2019), and to extract best practices and guidelines on the design of telemetry (Chung, 2015). The quality of the telemetry data and RCT design, coherent game design, and external measures have led to researchers continuing to use the CATS RCT dataset to develop and explore new methods for process data analysis (Feng & Cai, 2024, 2025; Tadayon & Pottie, 2020).

Advanced Statistical Modeling

A fourth innovation was the advancement of methodology relevant to large-scale educational effectiveness studies. Cai et al. (2016) developed a novel way to account for many of the constraints inherent in multisite RCT study designs. Using the CATS RCT data, Cai et al. accounted for the RCT design constraints by using a multilevel two-tier item factor model to model latent gain. Cai et al.'s method was more precise in estimating effectiveness by being able to isolate the part of the posttest variance that was sensitive to change. The resulting effect size of $d = 0.57$

was more than twice the magnitude of the effect size computed for CATS using a classical measurement approach (d = 0.23) and used by WWC in its reviews of educational intervention studies (ED, IES, WWC, 2015).

Game-Based Assessment

The potential of using games for assessment purposes has been of interest to the measurement and assessment communities for some time (for a discussion of these issues related to games, see Baker et al., 2011; Baker & Delacruz, 2008, 2016; Delacruz, 2011; DiCerbo et al., 2016; Landers, 2015; Mislevy et al., 2015; Oranje et al., 2019; OECD, 2014, 2021; Shute & Wang, 2016; Sireci, 2016; for a discussion of these issues related to process data in assessments, see Jiao et al., 2021; Lindner & Greiff, 2023; Zumbo et al., 2023). A common aspirational goal is to "replace the dull, time-consuming, and anxiety-producing traditional approaches commonly used today" (Landers, 2015, p. vii). Landers's sentiment reflects the general desire to develop other means of measuring what learners know and can do under more engaging and complex situations.

While there may be much interest in using games for assessment purposes, numerous literature reviews have found few studies that gathered validity evidence about how games in general and game mechanics in particular relate to knowledge, skills, and learning processes (Chung & Feng, 2024; see reviews by Gómez et al., 2022; Gris & Bengtson, 2021; Kim & Ifenthaler, 2019; Tlili et al., 2021; Wiley et al., 2021). In the remainder of this section, we describe some of the R&D related to gathering such validity evidence.

Identification of Game Features That Facilitate Measurement

One of the continuous efforts in CRESST's games-related R&D has been to identify game features to support measurement. The features were identified through usability studies, qualitative feature analysis, repeated observation of similar patterns of results, and data cleaning and algorithm development. A set of the most important features are described next.

When considering a game for measurement purposes, we think the most important game feature is the alignment among the game design, game mechanics, cognitive demands evoked by the game, and the external measure used to measure the learning outcomes of the game (Baker et al., 2011; Baker & Delacruz, 2008, 2016). For example, if a game is intended to promote computational thinking, then the

gameplay should require learners to engage in the critical computational thinking processes of designing a solution, failing, debugging, and iteration. A game that minimizes learner failures and errors will not be able to detect gaps in knowledge or the presence of misconceptions because players will have few opportunities to make mistakes.

The underlying idea is that the only possible observable behaviors are the interactions the game permits. If understanding the full range of learner performance is important, then having the complement of understanding—not understanding as exhibited by errors and misconceptions—is extremely valuable because measures of success and measures of failure can provide converging validity evidence. More generally, learners with higher domain knowledge should demonstrate more productive behaviors and fewer unproductive behaviors, and learners with lower domain knowledge should demonstrate the opposite relations. We have consistently observed these complementary relations when tasks are tightly aligned with the external measures of domain knowledge (e.g., Chung & Feng, 2024).

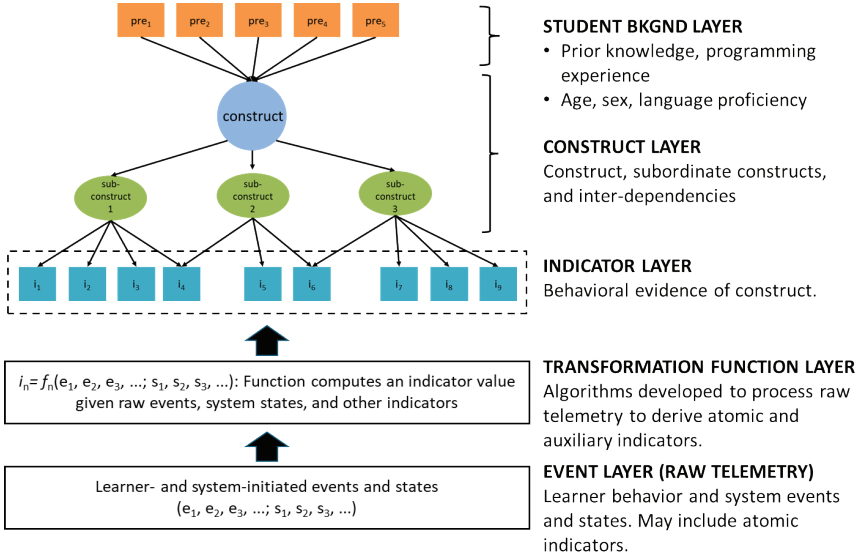
A second important game feature is practical. The user interface (UI) imposes constraints on learners' behavior (Chung & Baker, 2003). An important consideration is how to ensure that an action is intentional and not a mistake or other unwanted behavior that would contribute to construct-irrelevant variance. One type of UI element is the use of an explicit click (e.g., a button or similar UI element) that allows learners to signal, for example, that they are ready to move to the next level, to test a potential solution to a design, to select one option from a set of options, or to request help. Cleverly designed game mechanics can allow learners to perform such explicit actions as a natural part of the game. An explicit action also marks data and simplifies algorithm development by having explicit markers in the data to delineate time windows, sequences, and different levels of aggregation. Finally, game mechanics that require learners to render a judgment related to the content are especially useful if their choices can be evaluated (e.g., if moving a game piece can be evaluated as a correct or incorrect action).

Figure 5 shows how we think about fine-grained gameplay behavior (i.e., raw telemetry), indicators, and a measurement model. Indicator development often requires extensive data cleaning and processing to transform moment-to-moment events into inputs to statistical models. The programming task can range from

simply counting events to deriving numerous auxiliary variables to represent different game states that are themselves used to derive indicators. The encoding of useful information in the telemetry is dependent on both what the game allows learners to do through game mechanics, and the degree to which the game mechanics reflect the desired cognitive demands.

Figure 5.

Computational Modeling Conceptualization



Validity Evidence

Chung and Feng (2024) addressed the question, *To what extent do game-based indicators relate to criterion measures of learning*? drawing on various CRESST game-related studies. The authors reported that "common measures" composed of game performance and game progress indicators appear sensitive to the criterion measure across a broad set of games (See Chung & Feng, Appendix). The definition of game progress and game performance are game independent and analogous to the speed and accuracy variables studied extensively in verbal learning and motor learning. One use of game progress and game performance variables might also serve as a standardized metric to compare learning games

on their potential to promote knowledge or skill. See Chung and Feng (2024) and Chung and Roberts (2018) for additional examples.

The second type of indicators are game-specific indicators tailored to a game. For example, indicators of debugging behaviors were developed for a programming game (Feng & Chung, 2022), misconceptions developed for a pan balance game (Feng, 2019), deductive reasoning for a problem-solving game (Chung et al., 2018), and fractions misconceptions for a fractions game (Kerr, 2014). In all cases, the relation between the indicators and an external outcome measure were in the expected directions. Indicators that represent productive behaviors were often significantly and positively related to the external criterion measure, and indicators representing unproductive behavior were often significantly and *negatively* related to the external criterion measure (additional examples are presented in Choi, Parks, et al., 2021; Chung & Feng, 2024; Chung & Parks, 2015; Chung, Parks, et al., 2016; Redman et al., 2018, 2020, 2021, 2025; Roberts et al., 2016).

Application of Psychometric Modeling to Gameplay Data

One of the most important advances in game-based assessment was demonstrated by Feng and Cai (2024). In their study, the authors used the CATS RCT dataset to jointly model pretest, posttest, and gameplay data using a cross-classified IRT model. Feng and Cai modeled learners' latent changes in fractions knowledge and were able to directly relate the latent change to gameplay behavior. This new modeling approach directly provides information often of most interest in educational interventions: How much did learners learn (as described by latent changes in learners' knowledge over the course of instruction), and what variables influenced their learning (as described by learners' gameplay behavior)? Furthermore, the modeling technique is sufficiently general to incorporate other streams of data, such as multimodal data (e.g., eye tracking), learner background information, level design information, and interactions between learners' characteristics and the instructional setting.

Use of Population Data

One challenge presented by PBS KIDS (See Roberts et al., in press) was to examine how games played "in the wild" (i.e., the population) can be used to understand PBS KIDS' audience better. The only information available with population gameplay data is an anonymous ID. Three general issues were explored: using

psychometric modeling to estimate latent ability, using population-derived models and parameters in RCT studies, and testing a method to infer learning solely from players' gameplay behavior.

Psychometric Modeling of Population Gameplay Data

In numerous studies involving PBS KIDS' gameplay data from players "in the wild," CRESST applied various psychometric models. A close analysis of the game design and available gameplay indicators dictated the choice of models. The models included higher order IRT (de la Torre & Song, 2009) and diagnostic classification (Rupp et al., 2010) in Choi, Suh, et al. (2021); Rasch and Rasch Poisson counts (Rasch, 1960), IRT trees, and linear logistic testing model (De Boeck & Wilson, 2004) in Redman et al. (2021); a one-factor 2PL model, bifactor 2PL model with two and three specific factors in Redman et al. (2023); a multiple-group two-time point nominal IRT model (Cai, 2010; Cai & Houts, 2021) in Redman et al. (2025); and a two-time point graded response IRT model in Feng et al. (2025).

Using Population Information in RCT Studies.

To demonstrate how population data could be used in RCT studies, Choi, Parks, et al. (2021) used population gameplay to fit higher order IRT models for two PBS KIDS games. Choi, Suh, et al. (2021) used the population-based models and estimated model parameters from Choi, Parks, et al. (2021) to estimate ability of learners playing the same games in an RCT sample (Education Development Center, Inc., & SRI International, 2021). Diagnostic classification models (DCM) were also used to estimate informational text attribute profiles in both the RCT sample and population.

Estimating Learning in the Population Through Gameplay.

Finally, we explored the use of PBS KIDS games played "in the wild" to directly measure changes in gameplay that were consistent with changes in learning (Redman et al., 2023). The games were classified into three categories (likely, less likely, not likely) on their potential to promote learning. A two-timepoint latent variable model was used to estimate changes in latent ability using only game-based indicators. The study found that for the two games rated as not likely or less likely to result in learning, the effect sizes of the change in latent score were 0.07. In contrast, for the two games that were rated as likely to result in learning, the effect sizes of the change in latent score were 0.56 and 0.59.

Impact

The breadth of CRESST R&D around games for learning and games for assessment have led to insights about the conditions needed for both learning and measurement to be realized: Games that are effective in promoting learning can also yield information about learners' knowledge and skills, but only if (a) the game design and game mechanics in particular evoke the intended cognitive demands, (b) the game is instrumented to collect moment-to-moment telemetry and game state information, (c) the algorithms used to derive indicators from the telemetry are able to represent a range of performance, and (d) the psychometric models account for the constraints imposed by the game itself.

An important implication of this work for AISL is the idea of *measurement without testing*. Regardless of the type of task—game or otherwise—if the learner's behavior in the task is a manifestation of the desired cognitive demand, then the learner's behavior can serve as evidence of the cognitive demand occurring. This idea holds regardless of whether a task is designed for testing purposes or for learning purposes, for it is the interaction that is the atomic unit of observation.

Conclusion

This chapter presented a few examples of CRESST research extending over several years of effort and gave only a handful of references for each of them. Every area includes the importance of designing assessments to map to the purpose of evaluation and to provide as much transparency as possible. In most cases, our evaluations addressed not only performance on outcomes, but the value of instructional procedures and learner processes as well.

CRESST did not always juggle well the competing goals of innovation and early involvement with longer term impact. Much of our work was, in a self-aggrandizing sense, ahead of its time. This lack of fit with the context of learning and assessment vastly limited its immediate impact. However, we want to acknowledge and thank those educational and technology leaders who joined with us to explore learning and assessment strategies that were often too early for widespread use. There are numerous examples of other CRESST activities that affected proximal practice. The selection we chose to highlight, however, are focused on ideas that continue to affect educational research and development.

The methodologies and insights described in the examples also foreshadow the movement toward AISL, most clearly seen in the focus, since the inception of CRESST, on exploring assessment in the context of learning to support both attainment of learning goals and as an outcome measure. As the examples illustrate, designing assessments in the context of learning:

- Emphasizes measuring the most important concepts and skills.
- Conceives of human performance as being on a continuum, which naturally leads to the choice of experts as the criterion or reference against which to judge learner performance.
- Situates cognitive demands as a core assessment design requirement. By specifying and unpacking the key learning processes and outcomes a task is expected to evoke from learners, the assessment design process can be focused. Clear specifications can guide the development of measures, instructional content, and professional development.
- Treats the quality of measures as a necessary condition for drawing valid inferences by having clear and comprehensive definitions of what is to be measured, by making explicit how a student response is transformed into a quantitative value, and gathering validity evidence that the measures behave in expected ways.
- Is agnostic on the instructional or assessment setting, as well as the media, mode, and format used for instruction or assessment. Paper, digital, selected-response or constructed-response modes and formats can provide different information under different situations.
- Does not preclude a learning task from providing measurement information. A learning task can provide information about learners' ongoing knowledge and skills if learners are able to actually engage in the target cognitive demands and if learners' behaviors can be captured and stored.

As the assessment enterprise moves increasingly toward AISL, we think CRESST's experience can shed light on some of the challenges and opportunities ahead. The most important challenge is an understanding of cognitive demands and its implications for task design, the types and range of learner responses evoked by the task, and data capture opportunities. Additionally, adopting a naive view of measurement may be helpful for alignment, especially in technology-based

environments. If we think of the initial stages of measurement as simply an observation with some quantitative value assigned to it, then we can view a task as a set of learner-system interactions. Most of the interactions will be of little interest, but interactions that reflect judgment, decision making, or application of the target knowledge can be highly informative because they presumably reflect the outputs of learners' knowledge and skill. Furthermore, these interactions can be thought of as atomic units that can be combined, sequenced, or aggregated to form indicators that match future claims and inferences. Finally, this conceptualization, used in our work in simulations and games, can be applied to any environment where interactions exist. The limiting factor is observational capability.

The examples in this chapter addressed the Handbook principles of *transparency*, *purpose* and *focus*, and *validity.* As the field moves to more technology-based solutions, we think these principles become even more salient. Complex technology often obfuscates what is actually happening "under the hood" making independent inspection and critique nearly impossible. One path to make such systems more transparent is to develop tools and methods to specify in a formal way what to measure and the rules for transforming an observation into a measure. Another path is the training of assessment designers and technology developers on the AISL principles, methodologies, and insights described in this chapter so that best practices are designed into the applications. Regardless of approach, we are confident that AISL can be realized.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77,* 598–614.

Aleven, V., Dow, S., Christel, M., Stevens, S., Rosé, C., Koedinger, K., Myers, B., Flynn, J. B., Hintzman, Z., Harpstead, E., Hwang, S., Lomas, D., Reid, C., Yannier, N., Fathollahpour, M., Glenn, A., Sewall, J., Balash, J., Bastida, N., & Zhang, X. (2013). Supporting social-emotional development in collaborative inquiry games for K-3 science learning. In *Proceedings of the 9th Games+ Learning+ Society Conference-GLS* (Vol. 9, pp. 53–60). ETC Press.

Baker, E. L. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology, 14*(6), 10–16.

Baker, E. L. (1981, April 13–17). *Issues in the evaluation of composition instruction* [Paper presentation]. Annual meeting of the American Educational Research Association, Los Angeles, CA, United States.

Baker, E. L. (1982). The specification of writing tasks. *Evaluation in Education: An International Review Series*, *5*(3), 291–297.

Baker, E. L. (1988). Evaluating new technology: Formative evaluation of intelligent computer assisted instruction. In R. J. Seidel & P. D. Weddle (Eds.), *Computer-based instruction in military environments* (pp. 155–162). Plenum Publishing.

Baker, E. L. (1989, March 27–31). *The role of outcome measurement in the development and assessment of AI-based educational systems* [Paper presentation]. Annual meeting of the American Educational Research Association, San Francisco, CA, United States.

Baker, E. L. (1994). Human benchmarking of natural language systems. In H. F. O'Neil & E. L. Baker (Eds.), *Technology assessment of software applications* (pp. 85–98). Erlbaum.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247–254.

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment* (Special Issue), *12*(3&4), 179–194.

Baker, E. L. (2012). *Ontology-based educational design: Seeing is believing* (Resource Paper No. 13). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L. (2015). *Final report: Gamechanger: Using Technology to Improve Young Children's STEM Learning* (Deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Baron, J., Coffman, W., Freedman, S., Quellmalz, E., & Williams, P. (1986, September). *Issues in writing assessment for NAEP* [Paper presentation]. Office of Technology Assessment, Washington, DC.

Baker, E. L., & Butler, F. A. (1991). *Artificial intelligence measurement system: Overview and lessons learned* (ED332677). ERIC. https://files.eric.ed.gov/fulltext/ED332677.pdf

Baker, E. L., Choi, K., & O'Neil, H. F. (2022). The training assessment framework: Innovative tools using scenario-based assessment and feature analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, B. Goldberg, A. J. Hampton, & J. H. Johnston (Eds.), *Design recommendations for intelligent tutoring systems* (pp. 31–39). U.S. Army Combat Capabilities Development Command.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2011). The best and future uses of assessment in games. In *Technology-based assessments for 21st century skills* (pp. 227–246). Information Age Publishing.

Baker, E. L., Chung, G. K. W. K., Delacruz, G. C., & Griffin, N. C. (2012). *Engage validation plan* (Year 1 Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Clayton, S., Aschbacher, P., Chang, S., & Ni, Y. (1990, April 16–20). *Measuring deep understanding of history: The integration of prior knowledge and knowledge acquisition in explanations* [Paper presentation]. Annual meeting of the American Educational Research Association, Boston, MA, United States.

Baker, E. L., & Delacruz, G. C. (2008). A framework for the assessment of learning games. In H. F. O'Neil & R. S. Perez (Eds.), *Computer games and team and individual learning* (pp. 21–37). Elsevier.

Baker, E. L., & Delacruz, G. (2016). A framework to create effective learning games and simulations. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 3–20). Routledge/ Taylor & Francis.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Prentice-Hall.

Baker, E. L., & Gordon, E. W. (2014). From the assessment OF education to assessment FOR education: Policy and futures. *Teachers College Record*, 116(11), 1–24.

Baker, E. L., Koenig, A. D., Lee, J. J., Choi, K., O'Neil, H. F., Michiuye, J., K., & Griffin, N. (2025, January 4–7). *The Training Assessment Framework project: Flexible design from classroom to AI* [Paper presentation]. Annual Hawaii International Conference on Education, Honolulu, HI, United States.

Baker, E. L., Lee, J. J., Rivera, N. M., Choi, K., Bewley, W. L., Stripling, R., O'Neil, H. F. Jr., & Redman, E. (2015). *Detection and computational analysis of psychological signals: Evaluation of digital library and SimSensei for veteran use* (Final deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Lindheim, E. L., & Skrzypek, J. (1988). *Directly comparing computer and human performance in language understanding and visual reasoning* (CSE Report 288). University of California, Los Angeles, Center for the Study of Evaluation.

Baker, E. L., Madni, A., Choi, K., Kim, J., Redman, E. H., Delacruz, G. C., Chung, G. K. W. K., Griffin, N. C., & O'Neil, H. F. (2016). *ENGAGE2: Computer games for science learning at early grades: Evaluation, assessment, and neuro-sensing investigation* (Deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*(3–4), 269–282.

Baker, E. L., McGaw, B., & Sutherland, S. (2002). *Maintaining GCE A level standards*. Qualifications and Curriculum Authority.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, Z., Staley, L., Linn, R. L., & Rogosa, D. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior*, *18*(6), 609–622. https://doi.org/10.1016/S0747-5632(02)00019-5

Baker, E. L., & Quellmalz, E. (1986, March 13–15). *Initial results for the U.S. writing study* [Paper presentation]. Conference on College Communication and Composition, New Orleans, LA, United States.

Bewley, W. L., Chung, G. K. W. K., Delacruz, G. C., & Baker, E. L. (2009). Assessment models and tools for virtual environment training. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 300–313). Praeger Security International.

Burton, R. R., & Brown, J. S. (1979). An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies*, *11*(1), 5–24. https://doi.org/10.1016/S0020-7373(79)80003-6

Buschang, R. E., Kerr, D., & Chung, G. K. W. K. (2012). *Examining feedback in an instructional video game using process data and error analysis* (CRESST Report 817). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612.

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multi-dimensional item response theory. *Psychometrika, 86*, 754–777.

Cai, L., Choi, K., & Kuhfeld, M. (2016). On the role of multilevel item response models in multisite evaluation studies for serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 280–301). Routledge. https://doi.org/10.4324/9781315817767

Center for Advanced Technology in Schools. (2012). *CATS developed games* (CRESST Resource Report No. 15). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013a). *CATS knowledge and item specifications: Functions*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013b). *CATS knowledge and item specifications: Rational number equivalence*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013c). *CATS knowledge and item specifications: Solving equations*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chi, M. T., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Psychology Press.

Choi, K., Parks, C. B., Feng, T., Redman, E. J. K. H., & Chung, G. K. W. K. (2021). *Molly of Denali analytics validation study final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Choi, K., Suh, Y. S., Chung, G. K. W. K., Redman, E. J. K. H., Feng, T., & Parks, C. B. (2021). *A secondary analysis of the Molly of Denali RCT data: Examining the relationship among game-based indicators, video usage, and external outcomes using advanced psychometric modeling and population data* (Deliverable to EDC). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K. (2015). Guidelines for the design, implementation, and analysis of game telemetry. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 59–79). Springer.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1662

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., Wainess, R., & Bittick, S. J. (2010, April 30–May 4). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design* [Structured poster session]. Annual meeting of the American Educational Research Association, Denver, CO, United States.

Chung, G. K. W. K., Choi, K., Baker, E., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior*, *18*, 669–684.

Chung, G. K. W. K., Delacruz, G. C., de Vries, L. F., Kim, J.-O., Bewley, W. L., de Souza e Silva, A. A., Sylvester, R. M., & Baker, E. L. (2004). *Determinants of rifle marksmanship performance: Predicting shooting performance with advanced distributed learning assessments* (Deliverable to Office of Naval Research). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., Baker, E. L., Lee, J. J., & Osmundson, E. (2016). *Towards individualized instruction with technology-enabled tools and methods* (CRESST Report 854). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the I/ITSEC*, *25*, 1811–1822.

Chung, G. K. W. K., & Feng, T. (2024). From clicks to constructs: An examination of validity evidence of game-based indicators derived from theory. In M. Sahin & D. Ifenthaler (Eds.), *Assessment analytics in education. Advances in analytics for learning and teaching* (pp. 327–354). Springer. https://doi.org/10.1007/978-3-031-56365-2_17

Chung, G. K. W. K., Madni, A., & Baker, E. L. (2015). *EAITS test and evaluation final report* (Deliverable to Raytheon BBN). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Madni, A., Iseli, M., Koenig, A., & Baker, E. L. (2014). *CRESST Assessment report: Validation of knowledge probe methodology* (Deliverable to Raytheon BBN). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Michiuye, J. K., Brill, D. G., Sinha, R., Saadat, F., de Vries, L. F., Delacruz, G. C., Bewley, W. L., & Baker, E. L. (2002). *CRESST human performance knowledge mapping system* (Final deliverable to the Office of Naval Research). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Nagashima, S. O., Espinosa, P. D., Berka, C., & Baker, E. L. (2008). *An exploratory investigation of the effect of individualized computer-based instruction on rifle marksmanship performance and skill* (CRESST Tech. Rep. No. 754). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., O'Neil, H. F., Delacruz, G. C., & Bewley, W. L. (2005). The role of affect on novices' rifle marksmanship performance. *Educational Assessment*, *10*, 257–275.

Chung, G. K. W. K., & Parks, C. (2015). *Bundle 1 computational model–v1 (Measurement)* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Parks, C. B., Redman, E. J. K. H., Choi, K., Kim, J., Madni, A., & Baker, E. L. (2016). *PBS KIDS final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Redman, E. H., & Madni, A. (2018, January 4–7). *Computer-based assessment of nonroutine problem solving* [Presentation]. Sixteenth Annual Hawaii International Conference on Education, Honolulu, HI, United States.

Chung, G. K. W. K., & Roberts, J. (2018, April 13–17). Common learning analytics for learning games. In E. L. Baker (Chair), *Games and simulations: Learning analytics and metrics* [Symposium]. Annual meeting of the American Educational Research Association, New York, NY, United States.

Chung, G. K. W. K., Ruan, Z., & Redman, E. J. K. H. (2021, April 9–12). *A qualitative comparison of young children's performance on analogous digital and hands-on tasks: Assessment implications* [Paper presentation]. Annual meeting of the American Educational Research Association, Virtual Conference, United States.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer Science & Business Media.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620–639.

Delacruz, G. C. (2011). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking* (CRESST Report 796). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

DiCerbo, K. E., Mislevy, R. J., & Behrens, J. T. (2016). Inference in game-based assessment. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 253–279). Routledge. https://doi.org/10.4324/9781315817767

Education Development Center, Inc., & SRI International. (2021). *Mahsi'choo for the Info! Molly of Denali teaches children about informational text*. https://www.edc.org/sites/default/files/uploads/EDC-SRI-Mahsi-choo-Info-Summary.pdf

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725–747.

Espinosa, P. D., Nagashima, S. O., Chung, G. K. W. K., Parks, D., & Baker, E. L. (2009). *Development of sensor-based measures of rifle marksmanship skill and performance* (CRESST Tech. Rep. No. 756). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Feng, T. (2019, April 5–9). *Using game-based measures to assess children's scientific thinking about force* [Poster presentation]. Annual meeting of the American Educational Research Association, Toronto, Canada.

Feng, T., & Cai, L. (2024). Sensemaking of process data from evaluation studies of educational games: An application of cross-classified item response theory modeling. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12396

Feng, T., & Cai, L. (2025). Integrating data from multiple sources in evaluation studies of educational games: An application of cross-classified item response theory modeling. In J. L. Plass & X. Ochoa (Eds.), *Serious games* (Vol. 15259, pp. 70–76). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74138-8_6

Feng, T., & Chung, G. K. W. K. (2022, April 22–25). Extracting debugging indicators based on distance to solution in a block-based programming game. In G. K. W. K. Chung (Chair), *Game-based indicators of learning processes: Extraction methods, validity evidence, and applications* [Symposium]. Annual meeting of the American Educational Research Association, San Diego, CA, United States.

Feng, T., Chung, G. K. W. K., Choi, K., & Redman, E. J. K. H. (2025). *EDC SRI Wombats secondary analysis–Final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Gates, A. I. (1918). The abilities of an expert marksman tested in the psychological laboratory. *Journal of Applied Psychology, 2,* 1–14.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner, & A. L. Stevens (Eds.), *Mental models* (pp. 99–129). Psychology Press.

Gómez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies,* 1–16. https://doi.org/10.1109/TLT.2022.3226661

Gorman, T. P., Purves, A. C., & Degenhart, R. E. (1988). *The IEA study of written composition I: The international writing tasks and scoring scales.* Pergamon.

Graves, D. (1978). *Balance the basics: Let them write.* Ford Foundation.

Gris, G., & Bengtson, C. (2021). Assessment measures in game-based learning research: A systematic review. *International Journal of Serious Games*, *8*(1), Article 1. https://doi.org/10.17083/ijsg.v8i1.383

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies, 43*(5), 907–928. https://doi.org/10.1006/ijhc.1995.1081

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of US history knowledge. *Journal of Educational Research*, *89*(4), 206–218.

Herl, H. E., O'Neil Jr, H. F., Chung, G. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior*, *15*(3–4), 315–333.

Ihlenfeldt, S., Chung, G. K. W., K., Lyons, S., Lawson, J., & Redman, E. J. K. H. (2025). *Modeling the relationships among online Solitaire gameplay and measures of cognition* (CRESST Report 877). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Iseli, M. R., & Jha, R. (2016). Computational issues in modeling user behavior in serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 21–40). Routledge.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations* (CRESST Report 775). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Iseli, M. R., Lee, J. J., Schenke, K., Leon, S., Lim, D., Jones, B., & Cai, L. (2019). *Simulation-based assessment of ultrasound proficiency* (CRESST Report 865). UCLA/CRESST

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Editorial: Process data in educational and psychological measurement. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.793399

Kerr, D. (2014). *Identifying common mathematical misconceptions from actions in educational video games* (CRESST Report 838). UCLA/CRESST

Kerr, D., & Chung, G. K. W. K. (2012a). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*, 144–182.

Kerr, D., & Chung, G. K. W. K. (2012b). *Using cluster analysis to extend usability testing to instructional content* (CRESST Report 816). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kerr, D., & Chung, G. K. W. K. (2013a). Identifying learning trajectories in an educational video game. In R. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI Application Workshops: Big Data Meet Complex Models and Models for Spatial, Temporal and Network Data* (pp. 20–28). http://ceur-ws.org/Vol-1024/

Kerr, D., & Chung, G. K. W. K. (2013b). *The effect of in-game errors on learning outcomes* (CRESST Report 835). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisited: Advances in game-based learning* (pp. 3–11). Springer. https://doi.org/10.1007/978-3-030-15569-8_1

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation* (CRESST Report 771). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kumar, R., Chung, G. K. W. K., Madni, A., & Roberts, B. (2015). First evaluation of the physics instantiation of a problem-solving based online learning platform. In C. Conati, N. Hefferman, A. Mitrovic, & M. F. Verdejo (Eds.), *Lecture Notes in Computer Science: Vol. 9112. Artificial Intelligence in Education* (pp. 686–689). Springer.

Landers, R. (2015). Special issue on assessing human capabilities in video games and simulations. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), iv–viii.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, *54*(6), 771–794. https://doi.org/10.1080/00273171.2019.1590794

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment. *European Journal of Psychological Assessment*, *39*(4), 241–251. https://doi.org/10.1027/1015-5759/a000790

Madni, A., Griffin, N., & Yang, J. S. (2013, April 27–May 1). *Integrating assessment of SEL into an early childhood science learning context* [Paper presentation]. Annual meeting of the American Educational Research Association, San Francisco, CA, United States.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2015). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing* (pp. 23–48). Routledge. https://doi.org/10.4324/9781315871493

Nagashima, S. O., Chung, G. K. W. K., Espinosa, P. D., & Berka, C. (2009). Sensor-based assessment of basic rifle marksmanship. *Proceedings of the I/ITSEC*, Orlando, FL.

Niemi, D., & Baker, E. L. (1998, January). *Design and development of a comprehensive assessment system: Pilot testing, scoring, and refinement of mathematics and language arts performance assessments* (Final Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

O'Neil, H. F. Jr, & Baker, E. L. (1987). *Issues in intelligent computer-assisted instruction, evaluation and measurement*. In T. B. Gutkin & S. L. Wise (Eds.), *The computer and decision-making process* (pp. 199–224). Erlbaum.

O'Neil, H. F., Baker, E. L., & Linn, R. L. (1990, April 16–20). *Performance assessment framework* [Paper presentation]. Annual meeting of the American Educational Research Association, Boston, MA, United States.

O'Neil, H. F., Baker, E. L., Ni, Y., Jacoby, A., & Swigger, K. M. (1994). *Human benchmarking for the evaluation of expert systems.* In H. F. O'Neil, Jr., & E. L. Baker (Eds.), Technology assessment in software applications (pp. 13–45). Lawrence Erlbaum Associates.

O'Neil, H. F., Mayer, R. E., Rueda, R., & Baker, E. L. (2021). Measuring and increasing self-efficacy in a game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 131–158). Routledge/ Taylor & Francis.

Oranje, A., Mislevy, B., Bauer, M. I., & Jackson, G. T. (2019). Summative game-based assessment. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisited* (pp. 37–65). Springer. https://doi.org/10.1007/978-3-030-15569-8_3

Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Results: Creative problem solving: Students' skills in tackling real-life problems* (Volume V). OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en

Organisation for Economic Co-operation and Development (OECD). (2021). *OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain and robots*. https://doi.org/10.1787/589b283f-en

Quellmalz, E. S. (1982). *Designing writing assessments: Balancing fairness, utility, and cost* (CSE Report 188). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Parks, C. B., Schenke, K., Michiuye, J. K., Choi, K., Ziyue, R., & Wu, Z. (2020). *Cat in the Hat Builds That analytics validation study—final deliverable* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Chung, G. K. W. K., Schenke, K., Maierhofer, T., Parks, C. B., Chang, S. M., Feng, T., Riveroll, C. S., & Michiuye, J. K. (2018). *Connected learning final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Feng, T., Parks, C. B., Choi, K., & Chung, G. K. W. K. (2023). *Learning-related analytics KPI—KPI final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Feng, T., Parks, C. B., Chung, G. K. W. K., Choi, K., & Cai, L. (2025). *Wombats analytics evaluation –Final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Parks, C. B., Michiuye, J. K., Suh, Y. S., Chung, G. K. W. K., Kim, J., & Griffin, N. (2021). *Social-emotional learning games validity study (exploratory study): Final study report*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Richardson, M. W., Russell, J. T., Stalnaker, J. M., & Thurstone, L. L. (1933). *Manual of examination methods*. The University of Chicago. http://hdl.handle.net/2027/uiug.30112066775344

Roberts, J. D., Chung, G. K. W. K., & Parks, C. B. (2016). Supporting children's progress through the PBS KIDS learning analytics platform. *Journal of Children and Media, 10*, 257–266.

Roberts, J. D., Younger, J. W., Corrado, K., Felline, C., & Lovato, S. (in press). Practical examples of assessment in the service of learning at PBS KIDS. In Tucker, E., Everson, E., Baker, E. L., & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III, Examples of assessment in the service of learning*. University of Massachussets Amherst.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment*, *7*(2), 99–141.

Rupp, A. A., Templin J., & Henson R. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Savitsky, E. (2013). *U.S. Patent No. 8,480,404*. U.S. Patent and Trademark Office.

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science, 8*(2), 173–190.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior*, *15*, 403–418.

Shute, V., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment* (pp. 535–562). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch22

Sireci, S. G. (2016). Commentary on chapters 1–4: Using technology to enhance assessments. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 104–108). Routledge. https://doi.org/10.4324/9781315871493

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., & Clyman, S. (1999). Artificial neural network-based performance assessments. *Computers in Human Behavior*, *15*(3), 295–313. https://doi.org/10.1016/S0747-5632(99)00025-4

Tadayon, M., & Pottie, G. J. (2020). Predicting student performance in an educational game using a hidden Markov model. *IEEE Transactions on Education, 63*(4), 299–304.

Tlili, A., Chang, M., Moon, J., Liu, Z., Burgos, D., Chen, N.-S., & Kinshuk. (2021). A systematic literature review of empirical studies on learning analytics in educational games. *International Journal of Interactive Multimedia and Artificial Intelligence, 7*(2), 250–261. http://doi.org/10.9781/ijimai.2021.03.003

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2015, November). *WWC review of the report: The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques.* http://whatworks.ed.gov

U.S. Marine Corps. (2001). *Rifle marksmanship* (PCN 144 000091 00, MCRP 3–01A).

VanLehn, K., Chung, G., Grover, S., Madni, A., & Wetzel, J. (2016). Learning science by constructing models: Can Dragoon increase learning without increasing the time required? *International Journal of Artificial Intelligence in Education, 26*, 1033-1068. https://doi.org/10.1007/s40593-015-0093-5

Vendlinski, T. P., Chung, G. K. W. K., Binning, K. R., & Buschang, R. E. (2011). *Teaching rational number addition using video games: The effects of instructional variation.* (CRESST Report 808). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wigger, K. M., O'Neil, H. F., Jr., & Ni, Y. (1990). *Evaluation of expert systems: A review of the literature.* Los Angeles: University of California, Center for the Study of Evaluation/Center for Technology Assessment. https://apps.dtic.mil/sti/tr/pdf/ADA233601.pdf

Wiley, K., Robinson, R., & Mandryk, R. L. (2021). The making and evaluation of digital games used for the assessment of attention: Systematic review. *JMIR Serious Games, 9*(3), e26449. https://doi.org/10.2196/26449

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. *European Journal of Psychological Assessment*, *39*(4), 252–262. https://doi.org/10.1027/1015–5759/a00074.

Next Generation Science Standards: Challenges and Illustrations of Designing Assessments that Serve Learning

James W. Pellegrino and Howard T. Everson

Abstract

This chapter examines challenges and solutions in designing assessments aligned with the Next Generation Science Standards (NGSS), focusing on the NGSS's multi-dimensional approach to science education, integrating Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts. The chapter describes two major assessment design projects—the Next Generation Science Assessment (NGSA) project which developed classroom-focused assessment tasks for grades 3–8 that support formative assessment, and the Stackable, Instructionally-Embedded, Portable Science (SIPS) assessments which created end-of-unit assessments for grades 5 and 8. Both projects addressed the challenge of assessing integrated knowledge rather than separate dimensions of science learning. Throughout, the emphasis is on the importance of viewing science competence as a multi-dimensional performance that integrates content knowledge with scientific practices. The chapter concludes by discussing the benefits of these projects, including providing models for assessment design, creating ready-to-use resources for educators, and offering students challenging tasks that can better represent their scientific proficiency. While these efforts require further validation evidence with respect to their intended classroom use, the work described represents significant progress in developing assessments that align with contemporary views of science education while acknowledging the ongoing challenges in creating valid, reliable, and instructionally supportive measures of multi-dimensional science learning.

I. Changing Nature of Science Competence: What Students Need to Know and Be Able to Do

A. Multiple, Interconnected Dimensions of Competence

The nature of science competence has been reconsidered and the current conceptualization is most clearly expressed in the 2012 NRC report *A Framework for K–12 Science Education*, which articulates three interconnected dimensions of competence. The first of these dimensions are Disciplinary Core Ideas. In reaction to criticisms of U.S. science curricula being "a mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1997, p. 62) compared to other countries, the Framework identified and focused on a small set of core ideas in four areas: (a) life sciences, (b) physical sciences, (c) earth and space sciences, and (d) engineering, technology, and the application of science. In so doing, the Framework attempted to reduce the long and often disconnected catalog of factual knowledge that students typically had to memorize. Core ideas in the physical sciences include energy and matter, for example, and core ideas in the life sciences include ecosystems and biological evolution. Students are supposed to encounter these core ideas over the course of their school years at increasing levels of sophistication, deepening their knowledge over time. The second dimension is Crosscutting Concepts. The Framework identifies seven such concepts that have importance across many science disciplines; examples include patterns, cause and effect, systems thinking, and stability and change. The third dimension is Science and Engineering Practices. Eight key practices are identified, including asking questions (for science) and defining problems (for engineering); planning and carrying out investigations; developing and using models; analyzing and interpreting data, and engaging in argument from evidence.

While the Framework's three dimensions are conceptually distinct, the vision is one of coordination in science and engineering education such that the three are integrated in the teaching, learning, and doing of science and engineering. By engaging in the practices of science and engineering, students gain new knowledge about the disciplinary core ideas and come to understand the nature of how scientific knowledge develops. Thus, it is not just the description of key elements of each of the three dimensions that matters in defining science competence; the central argument of the Framework is that the meaning of competence is realized through performance expectations describing what students at various levels of educational experience should know and be able to do. These performance

expectations integrate the three dimensions and move beyond the vague terms, such as "know" and "understand," often used in previous science standards documents to more specific statements like ""analyze," "compare," "predict", and "model,"" in which the practices of science are wrapped around and integrated with core content. Finally, the Framework makes the case that competence and expertise develop over time and increase in sophistication and power as the product of coherent systems of curriculum, instruction, and assessment.

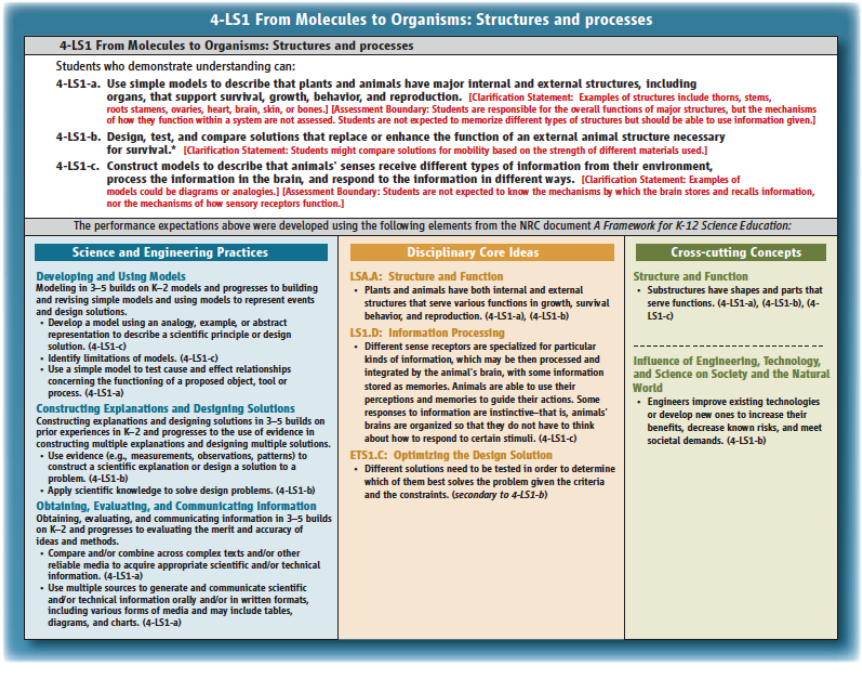
B. From Frameworks to Standards: A Focus on Performance Expectations

The Framework uses the three dimensions—the practices, crosscutting concepts, and core ideas of science and engineering—to organize the content and sequence of learning. This three-part structure signaled an important evolutionary shift for science education and presented the primary challenge for the design of both instruction and assessment—finding a way to describe and capture students' developing competence along these intertwined dimensions. The Framework emphasizes that research indicates that learning about science and engineering "involves integration of the knowledge of scientific explanations (i.e., content knowledge) and the practices needed to engage in scientific inquiry and engineering design" (p. 11). Both practices and crosscutting concepts are envisaged as tools (skills and strategies) for addressing new problems that are equally important for students' science learning as the domain knowledge topics with which they are integrated. Students who experience use of these tools in multiple contexts as they learn science are more likely to become flexible and effective users of them in new problem contexts.

To support the approach to science learning described above, the Framework states that assessment tasks must be designed to gather evidence of students' ability to apply the practices and their understanding of the crosscutting concepts in the contexts of problems that also require them to draw on their understanding of specific disciplinary ideas. In developing the Next Generation Science Standards (NGSS), Achieve and its partners elaborated these guidelines into standards that are clarified by descriptions of the ways in which students at each grade are expected to apply both the practices and crosscutting concepts, and of the knowledge they are expected to have of the core ideas (NGSS Lead States, 2013). As shown in Figure 1, the NGSS standards appear as clusters of performance expectations related to a particular aspect of a core disciplinary

idea. Each performance expectation asks students to use a specific practice and a crosscutting concept in the context of a specific element of the disciplinary knowledge relevant to a particular aspect of the core idea. Across the set of such expectations at a given grade level, each practice and crosscutting concept appears in multiple standards. Figure 1 shows the "architecture" of the performance expectations in terms of the underlying knowledge associated with each of the three facets of the Framework–disciplinary core ideas, science and engineering practices, and crosscutting concepts—for the set of three 4th grade performance expectations for the Life Science topic area labelled *From Molecules to Organisms: Structures and Processes*.

Figure 1.
Example of the NGSS Architecture for one Aspect of 4th grade Life Science.



In contrast to science standards like the NGSS that call for the integration of science practices and content knowledge, the prior generation of U.S. science standards (e.g., NRC, 1996) treated content and inquiry as fairly separate strands of science learning, and assessments followed suit. In some respects, the form the standards took contributed to this separation: content standards stated what students should know, and inquiry standards stated what they should be able to do. Consequently, assessments separately measured the knowledge and inquiry practice components. Thus, the idea of an integrated, multi-dimensional science performance presents a very different way of thinking about science proficiency. Disciplinary core ideas and crosscutting concepts serve as thinking tools that work together with scientific and engineering practices to enable learners to solve problems, reason with evidence, and make sense of phenomena. Such a view of competence also signifies that measuring proficiency solely as the acquisition of core content knowledge or as the ability to engage in inquiry processes free of content knowledge is neither appropriate nor sufficient.

C. Assessing Competence: How Will We Know What Students Know?

As illustrated in Figure 1, the NGSS performance expectations reflect intersections of a disciplinary core idea, science and engineering practices, and related crosscutting concepts, and they may also include boundary statements that identify limits to the level of understanding or context appropriate for a grade level and clarification statements that offer additional detail and examples. But standards and performance expectations, even as explicated in the NGSS, do not provide sufficient detail to create assessments. The design of valid and reliable science assessments is a complex endeavor that hinges on multiple elements that include, but are not restricted to, what is articulated in disciplinary frameworks and standards, such as those illustrated above for K–12 science education (Pellegrino et al., 2001; Mislevy & Haertel, 2006). For example, in the design of assessment items and tasks related to the performance expectations in Figure 1, one needs to also consider: (1) the kinds of conceptual models and evidence that we expect students to engage in; (2) grade-level appropriate contexts for assessing the performance expectations; (3) options for task design features (e.g., computer-based simulations, computer-based animations, paper-and-pencil writing and drawing) and which of these are essential for eliciting students' ideas about the performance expectation; and (4) the types of evidence that will reveal levels of student understanding and skill.

The challenge with standards expressed in this multi-dimensional form is how to design curricular and instructional materials to support acquisition of the important competencies underlying these performance expectations, and how to organize classroom instruction, including the design and use of formative and summative assessments, to promote student attainment of the complex disciplinary objectives embodied by such contemporary STEM standards. As discussed by Pellegrino, Wilson, Koenig, and Beatty in the 2014 NRC report *Developing Assessments for the Next Generation Science Standards*, significant assessment design challenges are posed by these multi-dimensional performance statements, especially when contrasted with previous generations of science assessment tasks that separately tested either disciplinary content knowledge or science "inquiry" (See also Pellegrino, 2013). They argued that considerable research and development was needed to create and evaluate assessment tasks and situations to determine if they can provide adequate and valid evidence of the proficiencies implied by the performance expectations of the NGSS, or any similar multi-dimensional standards derived from the NRC Framework.

Multiple arguments about the assessment design and validation challenges posed by the Framework and NGSS were explicated in some detail (Pellegrino et al., 2014), including the need for a principled design process to guide the work, of which the evidence centered design framework (Mislevy & Haertel, 2006) constitutes one such example. A related and critical argument was that such design and validation work needed to be conducted in instructional settings where students were being provided with adequate learning opportunities to construct the integrated knowledge envisioned by the NRC Framework and NGSS (Pellegrino, 2013; Pellegrino et al., 2014). While work of this type has advanced over the ensuing decade, much still needs to be done across the K–12 grade span and for multiple content domains. In the remainder of this chapter, we provide two examples of such efforts. Both focus on developing assessments and related instructional resources for use in K-8 classrooms. The two projects share an emphasis on supporting teachers as they strive to support students' progress toward developing and demonstrating the proficiencies underlying the performance expectations articulated in the Framework and NGSS. It is our contention that these two projects embody and support each of the multiple *Principles for Assessment in the Service of Learning* as espoused by Professor Edmund Gordon and his colleagues and as described in Volume I of this publication series.

II. The Next Generation Science Assessment (NGSA) Project

A. Introduction

As described above, the *Framework for K–12 Science Education* and the NGSS articulate an ambitious vision for what students should know and be able to do in science. They emphasize that all students must have the opportunity to learn and actively participate in authentic science through using and applying disciplinary core ideas (DCIs) in concert with science and engineering practices (SEPs) and crosscutting concepts (CCCs) to make sense of phenomena or solve problems. Central to this vision is the notion of knowledge-in-use, where students use and apply the three dimensions to build the integrated proficiencies identified in the NGSS Performance Expectations. Many science educators and scientists have embraced the vision described in the Framework and instantiated in the NGSS (e.g., NSTA, 2016), and the vast majority of states, representing more than 75% of the U.S. student population, now have standards influenced by the NGSS and/or the Framework. While this vision holds promise for engaging a broad diversity of students in the learning of science, the opportunity to learn can be realized only if teachers have the tools that can help them examine, reflect on, and improve their science instruction.

Among the most essential tools for teachers are classroom-based assessments. High-quality science instruction requires high-quality classroom-based assessments that can be used formatively and that are aligned with the standards (e.g., Fuhrman et al., 2009; Pellegrino et al., 2014; Pellegrino, 2018). Importantly, assessments provide a necessary picture of how students' science learning is building over time. Yet, many teachers do not feel well prepared to develop their own NGSS-aligned assessments or use them formatively in their classrooms (e.g., Furtak, 2017). Science teachers need purposefully designed assessment tasks for the NGSS that they can readily use in their classrooms. Especially needed are (1) tasks and rubrics that provide just-in-time information about students' progress in building toward the NGSS performance expectations (PEs), (2) resources that support instructional decision-making based on the assessment information, and (3) a delivery system for easy access and use by teachers and students.

The Next Generation Science Assessment project was initiated to address these needs by developing the NGSA System (http://nextgenscienceassessment.org). The system consists of innovative NGSS-aligned classroom-focused assessment tasks with rubrics for interpreting student performance and teacher guides for classroom use, all housed on an online portal for flexible administration and scoring

(https://ngss-assessment.portal.concord.org). As noted below, the NGSA System resources have been widely used both in the U.S. and internationally.

In the brief descriptions that follow we provide relevant background on the project's overall logic and need, the design team, the assessment design and development approach, validity evidence, and further information on the NGSA Portal's resources including some examples of resources.

B. Need for the NGSA System Resources

The NGSA Project Team pursued development of a technology-enabled assessment system for three important reasons. First, we know from considerable published literature and the wisdom of practice that assessment can be valuable for classroom pedagogy, especially when it is integrated within instruction and used formatively to guide the progress of student learning (e.g., Penuel & Shepard, 2016). But we also know that the NGSS Performance Expectations pose considerable challenges when it comes to designing assessments that support instruction and students' learning (Pellegrino et al., 2014). This creates a compelling reason to provide exemplar tasks and rubrics to teachers and others to illustrate what is expected of students and how to evaluate it.

Second, highly specified and developed resources (Cohen & Ball, 1999) are needed to help teachers integrate formative assessment practices into their instruction so that they can monitor students' progress. Indeed, well-designed assessment tasks are valuable for giving teachers a foothold to determine what their students know and can do—information that is also useful for making informed instructional decisions (Ruiz-Primo & Furtak, 2007; Ruiz-Primo & Furtak, 2024). However, assessment tasks alone are not enough. Enacting assessment tasks for formative use in classrooms presents unique problems of practice for teachers (Sezen-Barrie & Kelly, 2017), and these become even more pronounced when orchestrating science assessment within NGSS instruction (Furtak, 2017). Problems of practice include using tasks in formative ways and supporting students as they engage in tasks; interpreting student work; and determining next steps to advance student learning (e.g., Furtak, 2017; Kang, Thompson, & Windschitl, 2014; Shepard, Penuel, & Pellegrino, 2018). A viable solution is to provide teachers with assessment resources such as practice guides that illustrate how to formatively integrate assessment tasks into instruction over time, thereby making tasks usable and instructionally beneficial to teachers and their students.

Third, classroom assessments should take advantage of the capabilities provided by learning technologies. Technology-delivered assessments have several benefits for teachers and students to engage in regular formative assessment practice (Davies, 2010; Gane, Zaidi, & Pellegrino, 2018; Zhai & Wiebe, 2023). For students, technology enhancements such as video and simulations can expand the phenomena that can be investigated. Various assistive technologies can be used to make assessment materials more accessible to all students; for example, through screen readers that facilitate navigation and reading of text and speech-to-text capabilities that support students in responding to tasks. By providing background drawings, drawing tools, stamps, and/or predetermined model components, technologies can help scaffold students in demonstrating their learning in deeper ways. Moreover, because technology-delivered assessment tasks can enable students to use multiple modalities and representations, students with diverse abilities and language backgrounds may have better opportunities to demonstrate their proficiency than typical print-based assessments (Pellegrino & Quellmalz, 2010). For teachers, technology is well-suited to support implementation by providing scaffolding, data collection, and feedback features needed for effective formative use of assessment. Accordingly, technology-delivered assessments hold tremendous promise for supporting students in demonstrating their learning and for supporting teachers to implement assessments with relative ease and more readily interpret and use assessment information.

In summary, the NGSA project was designed to offer the field critical elements of a technology-supported comprehensive assessment system including a range of assessment tasks that can be used formatively to support science learning for all students.

C. The NGSA Design Team

The NGSA design and development team has been comprised of experts in science education, assessment, psychometrics, and technology from WestEd, the CREATE for STEM Institute at Michigan State University, the Learning Sciences Research Institute at the University of Illinois Chicago, and the Concord Consortium. This group initiated collaborative work in 2013, with an initial focus on developing NGSS-aligned assessment tasks and rubrics for instructionally supportive use in middle-school science classrooms. This was in response to the call for classroom focused assessment development and validation work in the NRC Report on

Developing Assessments for the NGSS (Pellegrino et al., 2014). Since the initial work on middle-school assessment, the collaborative has expanded to include experts from the STEM Education Center at the University of Chicago who have worked with other team members to develop assessment resources for upper elementary grades (3–5) teachers and students.

Across time, the group has worked closely with science teachers from multiple states and districts to develop usable and instructionally beneficial assessment tools that can help teachers better grasp the Framework and NGSS vision and more adeptly plan instruction to move students forward in their science learning. Final products developed by the team include teacher-tested and classroom-ready assessment tasks and rubrics that highlight learning in all three dimensions; guides to help teachers administer and interpret the assessment tasks and results; and an online platform that is searchable and enables teachers to assign tasks to students (individually or groups), monitor and obtain reports of student work, and access various support materials. The NGSA System is an open education resource housed in an online platform freely available to schools and districts with the explicit goal of promoting easy access and rapid adoption and use.

D. Development of the NGSA System's Resources

The current NGSA System was initially developed under the NSF-funded project, Collaborative Research: Designing Assessments in Physical Science Across Three Dimensions (DRL-1316903, 1903103, 1316908, & 1316874). In this project, the collaborative team developed a transformative approach for designing classroom-based assessment tasks that can provide teachers with meaningful and actionable information about students' progress toward achieving the NGSS PEs (See Harris, Krajcik, Pellegrino, & DeBarger, 2019). The approach follows the evidentiary reasoning logic of evidence-centered design (Mislevy & Haertel, 2006) and provides a systematic method for developing a variety of tasks that fulfill the important requirements for NGSS-designed assessment. Central to the design approach is the generation of sets of Learning Performances that establish targets to assess student progress towards mastery of the knowledge and competencies required by the PEs (Harris et al., 2018; McElhaney et al., 2016). The design approach is described in more detail in the following section.

The team used the design approach to iteratively develop tasks and rubrics aligned with a selected set of physical science PEs for the middle-school grade band. They also created the online task portal prototype through which the technology-based tasks could be delivered and used. In this initial work, the team also conducted task performance studies involving over 800 middle-school students (Gane et al., 2018) while also examining classroom use (Pennock & Severance, 2018; Zaidi et al., 2018; Gane et al., 2019). Subsequently, with funding support from the Gordon and Betty Moore Foundation and the Chan-Zuckerberg Initiative, the team completed the development of tasks and rubrics for all the physical science PEs. They also carried out early development work for some PEs in life science (tasks for four of the 21 life science PEs). All told, the team has produced an online bank of nearly 200 tasks designed to align with the middle-school PEs in the physical science and life science domains with accompanying resources. Most recently, with support from another NSF funded project–Collaborative Research: Improving Multi-dimensional Assessment and Instruction: Building and Sustaining Elementary Science Teachers' Capacity Through Learning Communities (Award #1813737 and #1813938), members of the NGSA team from UIC and STEM educators from the University of Chicago developed similar sets of resources for Performance Expectations spanning grades 3–5, including over 45 assessment tasks with accompanying rubrics and other resources.

E. Assessment Development: Design and Validation

NGSA Assessment Design Approach. The NGSA Project's approach to assessment design and development draws from evidence-centered design (ECD; Mislevy & Haertel, 2006). ECD emphasizes the evidentiary base for specifying coherent, logical relationships among the (a) learning goals that comprise the constructs to be measured (i.e., the claims articulating what students know and can do); (b) evidence in the form performances that should reveal the target constructs; and (c) features of tasks to elicit those performances. Using ECD, the design team created a principled approach for developing classroom-based science assessment of tasks that integrate the three dimensions (Harris et al., 2019). This approach allows for systematic derivation of a set of Learning Performances (LPs) from a single PE or bundle of PEs. LPs constitute knowledge-in-use statements that incorporate aspects of DCIs, SEPs, and CCCs that students need to be able to integrate as they progress toward achieving PEs. A single LP is smaller in scope and partially represents a PE. Taken collectively, a set of LPs describes the proficiencies that

students need to demonstrate to meet a PE. The project uses the LPs to guide the development of assessment tasks, evidence statements, and rubrics. Figure 2 presents a screenshot from the Portal showing the resources available to teachers for the Chemical Reactions topic area in middle school. Listed at the top are the three middle-school performance expectations that were bundled together under the Physical Science 1 middle-school topic area given their conceptual interrelationships to create the set of seven Learning Performances listed. Each of the seven Learning Performances covers a part of the conceptual space associated with the performance expectations for chemical reactions and each is stated as a three-dimensional expectation. Next to each Learning Performance is a button that expands to show the descriptions of two or more specific assessment tasks aligned to that specific Learning Performance. Teachers can then preview the sample tasks and find further information about them including rubrics that can be used for scoring student work.

Figure 2.
Illustration of Some Portal Resources for the Middle School Topic of Chemical Reactions.

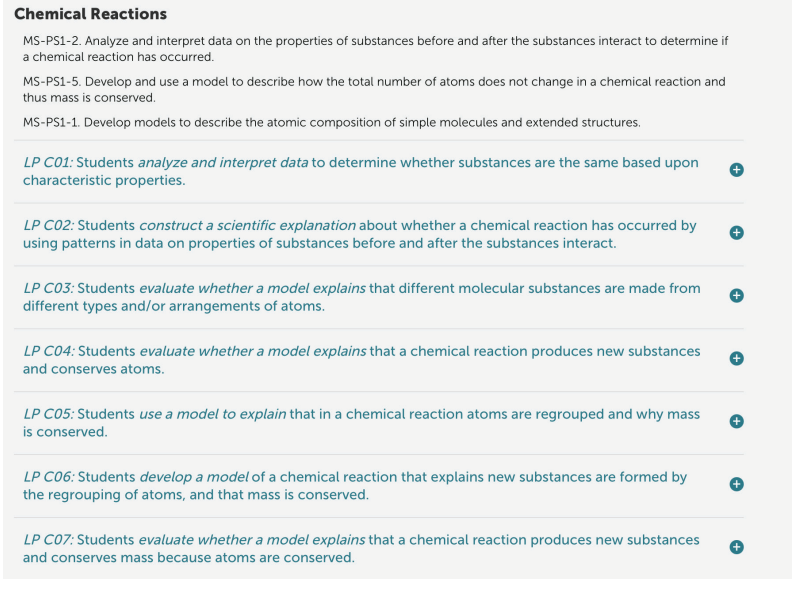
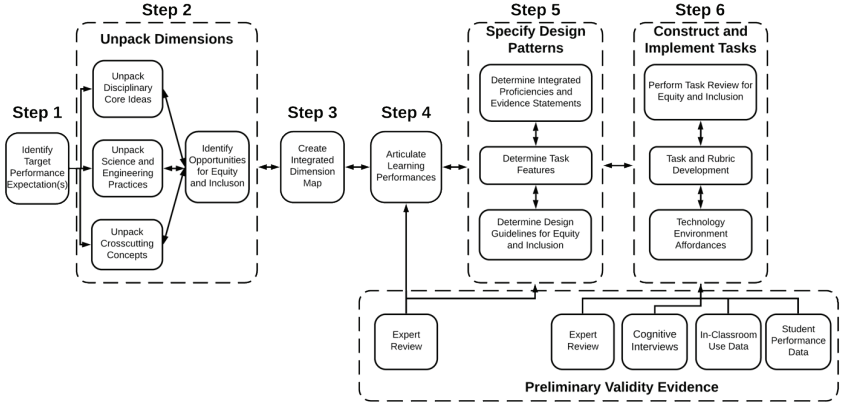


Figure 3 overviews the six-step design approach that was used to develop the actual tasks (for further information see Harris et al., 2019). Steps 1–3 are a domain analysis that entails unpacking the three NGSS dimensions of a PE(s). For the case illustrated In Figure 2, doing so involves consideration of the three PEs listed for chemical reactions. Unpacking the dimensions of the target PE(s) provides the anchors constituting each dimension and reveals a clear focus for what should be assessed. Integrated dimension maps are then created that provide a visual representation of the target PE(s). Steps 4 and 5 involve constructing Learning Performances such as those shown in Figure 2 and specifying design patterns for tasks associated with them. The integrated dimension map is used to articulate and refine a set of LPs that serve as claims, as they specify what students are expected to demonstrate for evidence that they have achieved one or more aspects of a PE. From each LP, design patterns are derived that include elements to ensure that the tasks elicit evidence of proficiency for the PE, notably evidence statements that articulate the observable features of student performance, equity and fairness considerations for characteristic task features, aspects common to all tasks, and variable task features, such as levels of scaffolding that vary from task to task. The final step in the design process, Step 6, involves using the design patterns to create tasks and accompanying rubrics.

Figure 3.
Overview of the NGSA Design Process



NGSA Validation Activities. In parallel with the design and development work, attention is given to the validation of the design products via multiple forms of evidence obtained during the design and implementation process as shown in Figure 3 (See Pellegrino et al., 2016). Detailed discussions of specific validation activities and results for the middle-school physical science and life science assessments can be found in several papers (e.g., Alozie et al., 2018; Gane et al., 2018, 2019; McElhaney et al., 2018; Zaidi et al., 2018).

Each stage in the process involves an independent review of products by science and science education experts. They review the integrated dimension maps, and the LPs derived from them. These same experts review the tasks designed to align with each LP and corresponding design pattern. Throughout the process we conduct an equity/fairness review to minimize bias. Once tasks have been through the expert review phases, they are further refined using several steps, including cognitive interviews with students that examine whether tasks are comprehensible and whether they elicit the target performance, collection of classroom performance data to determine applicability and reliability of scoring rules using the rubrics, and classroom studies with teachers who provide design feedback on tasks and help us consider strategies for formative use.

Equity and Inclusion are critical elements that are woven throughout the design and validation process, beginning with (a) the initial domain analysis of the PEs, and continuing through (b) the development of tasks, rubrics, and teacher guides; (c) recruitment of teacher and student participants; and (d) data analyses for validation. Moreover, by conducting the development work with teachers in districts across states that have adopted the NGSS, each serving distinct student populations, the project has been able to further ensure that the tasks and overall system are usable in diverse classroom settings and for broad access and participation.

F. Key Features of the NGSA System

As noted earlier, the NGSA System consists of a library of NGSS-designed tasks, teacher resources for implementing a formative assessment approach, and an online platform for task delivery and access to resources. What follows is some further information on the tasks, the teacher resources, and the open access portal.

NGSS-designed assessment tasks and teacher resources. Each task, anchored in a phenomenon and contextualized within a brief scenario, requires anywhere from 5 to 15 minutes to complete, depending on the requirements of the task. The shorter task duration balances the desire to engage students in authentic science practices with the need for teachers to use the tasks flexibly during instruction and to get timely information from the tasks for formative purposes. Because the task authoring system is web-based it is possible to integrate computational models, which students can manipulate to explore phenomena and generate data. Videos of phenomena, a drawing tool, a system modeling tool, and data analysis tools are also embedded in tasks, providing innovative ways for students to use and apply SEPs, DCIs, and CCCs.

The resources available to teachers include scoring rubrics for pinpointing areas for student feedback and instructional support, strategies for effectively using the assessment tasks in classrooms, and practical guidance for using the NGSA online system. Accompanying each task is a rubric that differentiates levels of proficiency and that includes exemplar responses.

Figure 4 provides an example of a life science task that involves a model for an experiment related to photosynthesis. The middle-school performance expectation is MS-LS1–6. Construct a scientific explanation based on evidence for the role of photosynthesis in the cycling of matter and flow of energy into and out of organisms. The related Learning Performance is Students evaluate how well a model shows that plants and other photosynthetic organisms use energy from the Sun to drive the production of food (sugar) and oxygen.

Figure 4.
Illustrative Task Related to the Topic of Photosynthesis

Carmen's leaf model (ID #090-04-p04)

Tap text to listen

Carmen's class is growing plants. The class wanted to investigate the role of oxygen and sugar in plants. They followed the steps below.

1. On each plant, they chose one leaf and covered it up halfway with foil to block the light it receives. The rest of the leaves on each plant were not covered with foil.
2. They then placed the plants in the sunlight for 1 day.
3. After a full day of sunlight, the foil was removed from the partially covered leaves.
4. All the leaves of the plant were soaked in iodine. **Iodine shows that sugar molecules are stored and will turn a purplish-black color.**
5. The plant leaves were observed and compared.

Results of Experiment: The image below shows the results of the experiment.

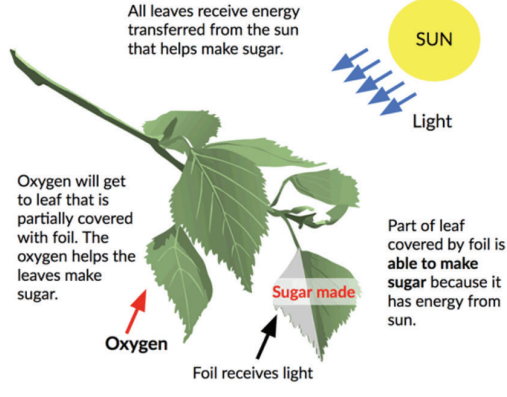
The model that Carmen and her classmates came up with to show their results is shown to the right.

Question #1

Carmen thinks there are errors in how the model explains what happens with sugar and oxygen on the half of the leaf that **was** covered with foil.

Desc Interactive content n the model. Explain how Carmen should improve the model In your response, use what you know about sunlight and oxygen and sugar in plants.

Please type your answer here.



Task Portal. The online portal (https://ngss-assessment.portal.concord.org) houses the current task library and teacher resources and includes a range of features for practitioners and researchers. Teachers can set up classes, assign tasks, receive reports of student work, and gain access to the resources linked to each task. As students work through tasks, their progress can be monitored in real-time. Teachers can review student responses and provide feedback via the portal using rubric-based responses, written notes, or scores. The portal also supports research activities, allowing tasks to be earmarked for research use and can even be tagged for specific research cohort designations.

The NGSA System's assessment tasks and supporting instructional resources for elementary and middle school have been in use in classrooms around the U.S. for several years. The online portal currently has more than 11,000 registered teacher accounts and over 85,000 registered student accounts. Registering an account enables teachers to directly assign tasks to students, access teacher guides, and collect and organize student work. However, to make it convenient for users, the use of the portal and its tasks alone does not require registration, so there is also a substantial "unregistered" user base. Overall, most users are from the U.S., with participation from every state, as well as some international interest with visitors from 126 countries. The user base continues to grow and team members are contacted regularly by teachers and districts with requests to expand the task library to include tasks covering more of the NGSS' elementary and middle grade PEs.

In addition to all the resources contained on the Portal, the team has published a book that serves as a guide for teachers and other educators to develop and use the design process to create similar types of tasks for use in their own classrooms. The volume is published by NSTA Press and titled Creating and Using Instructionally Supportive Assessments in NGSS Classrooms (Harris, Krajcik, & Pellegrino, 2024). Finally, the NGSA team has developed an open access website designed to support an ongoing Virtual Learning Community (VLC) for educators interested in the design and use of science assessments for classroom formative use. (https://www.upinscience.org). The VLC contains a variety of resources related to the formative assessment process and the use and interpretation of some of the tasks currently found on the Portal.

III. The Stackable, Instructionally-Embedded, Portable Science (SIPS) Assessments Project

In this section we review the rationale and goals of the SIPS project (hereafter the Project) and provide a brief summary of the pilot study that was conducted to test out key ideas for designs for assessing science learning in middle school as discussed in earlier Sections of this paper. We begin by describing the overall design thinking that guided the Project with selected illustrations, and then describe in broad strokes the multi-state pilot study we implemented to demonstrate a proof of concept that end-of-unit assessments could be developed and used by science teachers in their classrooms.

A. Rationale and Goals of the SIPS Project

As noted earlier, release of the NRC Framework and the NGSS standards shifted the focus to emphasize how well students can apply their science knowledge and this in turn has major implications for how assessments should be designed and developed to assess students' science learning (Pellegrino, 2013; Pellegrino et al., 2014). The Project was funded by the US Department of Education under the Competitive Grants for State Assessments Program, CFDA 84.368A. It brought together six states, five educational research organizations, and a panel of experts to address states' growing need for large-scale science assessments, as well as the needs of educators, parents, and students for resources that could support science learning throughout the school year. To meet this challenge the Project set out to build a bank of innovative science assessment tasks designed to measure students' learning that were carefully aligned with curricular and instructional resources to support ongoing instruction over the course of a school year. The term stackable in the Project's title indicates that the assessments can be used together sequentially or in varying orders across the academic year depending on the varying structure and sequence of local science instruction. They were designed to be embedded in the flow of instruction across the year with administration of the assessments proximal to the completion of each of a set of coherent instructional units. And they are portable because they can be used with a variety of science curricula and in a variety of instructional settings in and out of the classroom. The Project focused on grades five and eight as a proof of concept because these are the grades most often targeted in statewide science assessment systems.

To carry out the Project's research and development plan, a collaboration of educational researchers and representatives from departments of elementary and secondary education from six states was organized to carry out the Project. The six states included Nebraska, Alabama, Alaska, Montana, New York, and Wyoming. The educational research team included learning scientists, curriculum and instruction experts, assessment designers, and measurement experts from edCount LLC, the Learning Sciences Research Institute (LSRI) at the University of Illinois Chicago, SRI International, the National Center for the Improvement of Educational Assessment, and the Creative Measurement Solutions group.

B. Approach to Curriculum-Instruction-Assessment Design

The design team was charged with producing a wide range of science assessment resources for public access and use that are coordinated and aligned across all parts of a standards-based system for teaching and learning science that emphasized the interplay of curriculum, instruction, and assessment. The Project was grounded by the idea that to achieve coherence, the Curriculum-Assessment-Instruction (Pellegrino, 2010) connections ought to be balanced among our expectations and plans for student learning, how we carry out science instruction in classrooms, and how we assess students' science learning. With coherence as the guiding principle, the Project identified meaningful bundles of Next Generation Science Standards (NGSS) performance expectations for both grades 5 and 8 and created four instructional unit maps (i.e., instructional frameworks) that covered those expectations. An eighth-grade unit bundle of performance expectations for Force and Energy for grade 8 is shown in Figure 5.

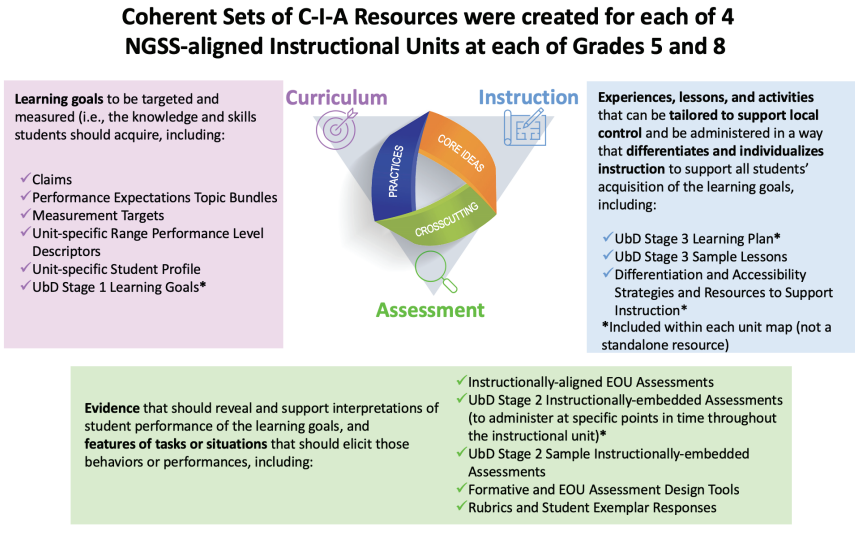
Figure 5.
Eighth Grade Unit Bundle of Performance Expectations

NGSS Grade 8 Unit 1: Forces and Energy
Bundle 1
MS-PS2–2. Plan an investigation to provide evidence that the change in an object's motion depends on the sum of the forces on the object and the mass of the object. **MS-PS2–1.** Apply Newton's Third Law to design a solution to a problem involving the motion of two colliding objects. **MS-PS3–1.** Construct and interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object. **MS-PS2–4.** Construct and present arguments using evidence to support the claim that gravitational interactions are attractive and depend on the masses of interacting objects.

For each unit, a unit map was created, and it encompassed a suite of interconnected and coherent curriculum, instruction, and assessment resources, all designed to support high-quality, three-dimensional science teaching and learning along a year-long instructional pathway. Figure 6 provides an overview of the design logic and lists the design elements and products generated under each of the Curriculum, Instruction, and Assessment components of the Unit design process. Figure 7 provides an illustration of the specific sets of resources created for the eighth-grade unit on Forces and Energy. Similar resources were created for all four eighth-grade units and all four fifth-grade units. All resources for each unit at each grade level can be accessed at the SIP Project website. (https://sipsassessments.org/resources/).

Figure 6.

Overview of the Sets of Resources Created for Each Instructional Unit.



https://sipsassessments.org/resources/

Figure 7.

Illustration of the Resources Created and Available for the 8th Grade Unit on Forces and Energy.

Grade 8 Unit 1: Forces and Energy

The Grade 8 Unit 1 topic, "Forces and Energy," organizes the Next Generation Science Standards performance expectations with a focus on helping students develop an understanding of the motion of objects and how interactions between objects can be explained and predicted.

Grade 8 Unit 1 Curriculum, Instruction, and Assessment Resources

Unit Map / Instructional Framework (.docx, .pdf)

Curriculum

- Claim, Measurement Target, and PE Bundle (.docx, .pdf)
- Storyline Overview (.pptx, .pdf)
- Student Profile (.docx, .pdf)
- Policy and Range Performance Level Descriptors (.docx, .pdf)
- Stage 1 Learning Goals* (See Unit Map / Instructional Framework)

Assessment

- Stage 2 Instructionally-embedded Assessments* (See Unit Map / Instructional Framework)
- Designing Equitable Assessments for Diverse Learners (.docx, .pdf)
- Sample Instructionally-embedded Assessments:
  - Segment 3: "Kinetic Energy vs. Mass/Speed Investigation"
    - Task Specification Tool (.docx, .pdf)
    - Task (.docx, .pdf)
  - Segment 4: "Designing Solutions to a Problem Involving a Collision"
    - Task Specification Tool (.docx, .pdf)
    - Task (.docx, .pdf)
- End-of-Unit Assessment (.docx, .pdf)
  - Assessment Scoring Guide (.docx, .pdf)
  - Design Tools:
    - Unpacking Tool (.docx, .pdf)
    - Design Pattern (.docx, .pdf)
    - Task 1 Specification Tool: "Storing Grocery Carts" (.docx, .pdf)
    - Task 2 Specification Tool: "Barriers on the Highway" (.docx, .pdf)
    - Task 3 Specification Tool: "Roller Coaster Thrills" (.docx, .pdf)

Instruction

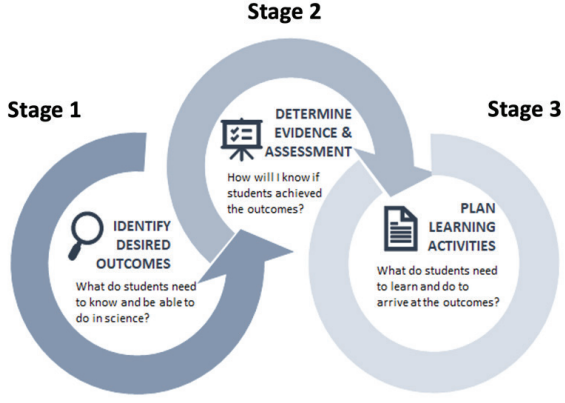
- Stage 3 Learning Plan* (See Unit Map / Instructional Framework)
- Differentiation Strategies and Resources (.docx, .pdf)
- Sample Lessons:
  - Segment 1: "Newton's Third Law" (.docx, .pdf)
  - Segment 2: "Getting to the Bottom of Newton's Second Law" (.docx, .pdf)

*Embedded within Unit Map / Instructional Framework

To move forward with this integrated design framework, the Project team drew on two heretofore and largely distinct approaches—a curriculum and instruction development approach known as Understanding by Design (UbD) (Wiggins & McTIghe, 2005) and the principled assessment design framework called Evidence Centered Design (ECD) discussed earlier and developed by Robert Mislevy and his colleagues (e.g., Mislevy, Haertel, Riconscente, Rutstein & Ziker, 2017).

Understanding by Design (UbD). The Project partners developed a prototype science curriculum framework based on the Understanding by Design (UbD) model of curriculum design. UbD uses a multi-stage method of backward planning that begins with a statement or vision of the desired results—the learning goals—and works backward to identify the assessment evidence needed to support inferences of student learning (See Figure 8). UbD calls for careful planning of the curriculum sequence and pedagogical tools and activities to achieve those stated learning goals. The UbD approach ensures that teachers are deliberately planning their lessons with a focus on the expected learning objectives and performance expectations of each of the science instructional units. Furthermore, UbD provides a framework for aligning the assessment design with the taught curriculum and the sources of evidence of student learning. A more complete description of UbD is beyond the scope of this chapter and the interested reader can find a richer description of this approach in Wiggins and McTighe, 2005.

Figure 8.
Simplified Representation of the three Stages of the Understanding by Design Framework



Source: Adapted from Wiggins, G.P. & McTighe, J. (2005).

Evidence Centered Design (ECD) and End-of-Unit Assessments. To design end-of-unit (EOU) assessments in a way that ensures alignment with the curricular frameworks and the relevant instructional resources the design team adapted a principled assessment design approach, i.e., ECD, to design and develop each of the Grade 5 and Grade 8 assessments (Mislevy & Haertel, 2006; Mislevy, Haertel, Riconscente, Rutstein & Ziker, 2017). Like the approach described earlier for the NGSA project, the team addressed these three key design questions: 1) what constructs do we want to measure; 2) what evidence is needed to make inferences about students' ability related to those constructs; and 3) how can tasks be designed to collect the desired evidence? Other explicit design criteria included the need to administer the EOUs at the end of completion of each of four instructional units—approximately every 10–12 weeks of science instruction; and they had to be administered by teachers within one 50-minute class session. Again, a more detailed description of the ECD methodology is beyond the scope of this chapter. The interested reader can find more thorough descriptions of this approach in the early work of Mislevy and Haertel (2006) and Mislevy & Riconscente (2006).

The ECD approach led us to compose each EOU assessment as a set of three sub-tasks, each containing multiple prompts (i.e., test items). The component tasks were designed to measure well-defined science constructs based on a clearly articulated theory of science learning. The aim was that any given assessment would produce evidence of students' science learning in terms of the NGSS performance expectations (PEs) that were the focus of the associated instructional unit. They were meant to provide a summative characterization of student learning as an outcome of the immediate prior instructional unit, as well as to inform the content and focus of subsequent instructional units. The evidence produced by the EOUs, by design and following the NGSA system described earlier, would support inferences about students' proficiency in integrating Scientific and Engineering Practices (SEPs) with important Disciplinary Core Ideas (DCIs) and Cross-cutting Concepts (CCCs) to scientifically investigate and understand natural phenomena and solve important science and engineering design problems. To make the multi-dimensional assessment design feasible, the design team defined proficiency and determined bundles of PEs that could be taught and measured together and that would meaningfully represent the scope of an instructional unit.

Each EOU assessment measured the key knowledge, skills, and abilities (the KSAs) as represented by a thorough unpacking of the PEs within the associated

instructional unit bundle identified during the UbD analysis process. Each PE was a combination of three dimensions: the disciplinary core ideas (DCI), science and engineering practices (SEPs), and cross-cutting concepts (CCC). Each of these dimensions was not unique to a given PE (e.g., the same scientific practice appears in multiple PEs), but the PE uniquely defines one combination of the three dimensions.

Another key step in the process required the design team to collaborate with the science teachers to develop a set of performance level descriptors (PLDs). These descriptors organized multi-dimensional statements into levels representing different levels of student performance. The PLDs provided statements that are at a finer grain size than the overall claim and provided further insight into what is to be measured on the assessment. Once the PLDs were developed, the design team created task design patterns for each PE in the instructional unit bundle.

In practice the design patterns provided task designers with a menu of options to use when designing tasks aligned to the PEs. The design patterns and PLD documents provided guidance on what should be measured, as the PLD statements and the KSAs describe the measured concepts related to the bundle of PEs. The design patterns also provided information on what evidence is needed to measure these concepts (through the demonstration of learning). Once the design team established the design patterns, the next step was to determine how to measure these concepts.

Like all educational assessments, the assessments developed in this Project had constraints on their design; specifically, they needed to be able to be completed in approximately one class period, and they needed to be administered as paper/pencil tasks. With these constraints in mind, each EOU assessment consisted of three tasks, each using one scenario and/or phenomenon, and a set of questions related to that phenomenon. Another critical design feature for measuring three-dimensional science standards is to engage students in a chain of sense-making. Therefore, the set of prompts within each task required students to engage with different aspects of the scenario and meet the expectation of increasing the complexity of the required response. The design team anticipated that each individual task would take students 10 to 15 minutes to complete, and consequently, determined that each EOU assessment would consist of three tasks.

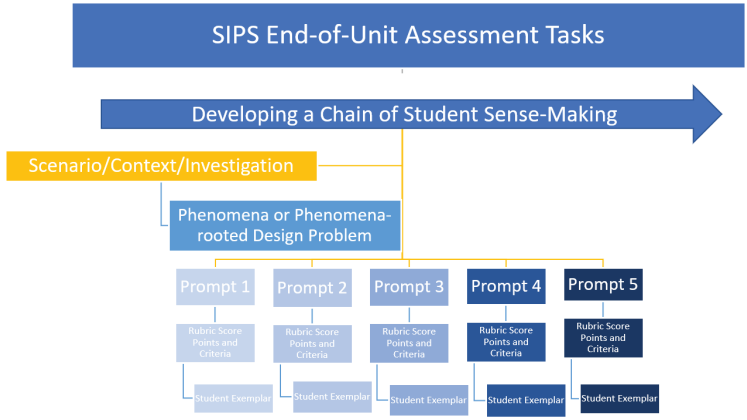
As noted previously, each EOU assessment consisted of three tasks. To provide further specifications for each task as part of an ECD approach, the design team created task specifications. Each task specification tool provides specification for the following:

- List of performance expectations covered in the task (each task covers one to two PEs);
- Information on the phenomenon or phenomenon-rooted design problem: Each task is rooted in a phenomenon or design problem related to the PEs;
- Scenario: Each task requires a scenario or situation which would make sense to students, be coherent and understandable to students, and provide enough context to allow students to engage meaningfully with the task;
- Variable Features: A list of features (or decision points) that could be modified to shift the complexity and/or focus of the task while still measuring the PEs;
- Chain of Sensemaking: An overview of the flow of the task, including the alignment of different sections to the KSAs;
- KSAs: A list of the KSAs that are targeted by the task, including any additional (not from the original set of design patterns) KSAs that are a cross between two PEs;
- Student Demonstration of Learning: A list of the expectations of students taken from the design patterns;
- Work Products: A list of the physical responses that students might produce;
- Application of Universal Design for Learning-based Guidelines: A set of guidelines to promote equity and inclusion in the task design; and
- SIPS Complexity Framework Components: A description of how the prompts for the task are designed to align with the degrees of sophistication represented by the complexity framework.

The task specification tool described the design elements of the task and provided guidance to task developers. This information was used to further develop the tasks. Each task is aligned to one or two PEs and is situated in a given phenomenon or design problem. The phenomenon was situated in an overall scenario and scaffolded such that students were provided a foundational context, the context is then problematized, and then students engage with the context through a series

of prompts or questions. The scenario had to make sense to students, be coherent and understandable, and provide enough context to allow students to engage meaningfully with the task. Again, leaning on the UbD approach, each task included rubrics that clearly defined what was required of students and how evidence from students could be evaluated. Figure 9, below, shows the components of an EOU assessment task.

Figure 9.
Illustration of the Components of an EOU Assessment Task



The EOU development process described above was used to produce eight prototype EOU assessments—four each at grade 5 and grade 8, all of which were intended to be administered after approximately 8 to 10 weeks of instruction (i.e., following each of the SIPS instructional units in each grade). Each assessment contains three multi-part tasks which are scenario/phenomena based and are designed in a way that students engage with sense-making as they move through the task.

To the extent possible, the task scenarios were based on a phenomenon or design problem that occurred outside of the classroom and has local or global relevance. However, given variation in curricular and instructional resources used across states and districts, SIPS partners acknowledge that tasks address phenomena

or phenomena-rooted design problems that may or may not have been addressed through instruction.

The tasks designed for each EOU were meant to be illustrative examples of (1) PE bundles and (2) task scenarios. Additional tasks can be designed using the SIPS design process to support use with other SIPS unit sequences or other curricula. While the EOUs were designed to be administered in the recommended order of the SIPS instructional units, if educators taught the instructional units in a different order then the assessments may be administered in the sequence that best aligns with instruction. Scoring for these assessments would be the same regardless of the order in which they are administered.

While not every prompt had to cover every dimension in the PE cluster, every dimension within the unit's PE bundle had to be aligned to at least one item on one task on the EOU assessment. Once tasks were developed, the design team reviewed the tasks for alignment against the task specification tool, ensuring coverage of the KSAs specified in the tool. Tasks were also reviewed for clarity, sense-making, accessibility and fairness, and the degree to which they require sense-making. Feedback was obtained from teachers as well as from outside experts and included reviews of the tasks as well as the scoring rubrics (described below). The Project design team applied revisions to the tasks based on this feedback.

Rubric Development. Scoring rubrics for each task were developed in conjunction with our science teacher partners to highlight aspects of the student responses that demonstrate understanding of the concepts. The scoring rubrics included evaluative criteria to support the evaluation of evidence for each prompt (or a set of sub-prompts) within each task and were developed based on the student demonstration of learning from the task specification tool. The number of score points possible for each prompt or set of sub-prompts varied from one to four points depending on the expectations of students.

Rubrics were designed with the expectation that teachers would be the primary users of the rubrics. Each score point was defined to provide clear guidelines of the differences between student responses that fall in each score point. Rubrics also cover the range of possible student responses and are specific to the given prompts as this allows for more guidance for scorers. Once the rubrics and tasks

were developed, the SIPS team aligned them back to the PLD descriptors, ensuring that the tasks and rubrics are focused on aspects of the PLDs that are deemed important and that the set of tasks as a whole cover the critical aspects of the PLDs. The SIPS team applied revisions to either the tasks or the PLDs (as concepts of the PLDs changed throughout the development process).

C. Pilot Study Overview and Results

To collect evidence about the validity and utility of the EOU assessments, a small pilot study was designed to focus on three overarching research questions: (1) to what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?: (2) how well do latent variable measurement models fit the empirical EOU assessment data?; and (3) overall, what do the EOU assessment results tell us about students' science learning? To address these issues, we recruited at least five classrooms of students from each state—aiming for a mix of grade 5 and grade 8 classrooms. See Table 1 for an overview of the teachers and students who participated in the pilot study.

Table 1.
Number of Educators and Students Included in the Pilot Study by EOU Assessment

EOU Assessment	Number of Teachers	Number of Students
Grade 5 Unit 1	23	341
Grade 5 Unit 2	28	473
Grade 5 Unit 3	19	341
Grade 5 Unit 4	26	417
Grade 8 Unit 1	14	151
Grade 8 Unit 2	10	189
Grade 8 Unit 3	13	258
Grade 8 Unit 4	4	51

The main requirement for educators to participate was teaching a curriculum aligned to three-dimensional science standards (e.g., NGSS standards or similar). In the end, the Project recruited 121 educators from across four states that expressed initial interest in participating in the pilot. Of those 121 educators, 63 educators representing three of the six partner states participated in the study by administering one or more EOU assessments.

Summary of Findings from Pilot Study. It is important to note that the study was designed as a pilot of a limited set of initial prototypes of each of the four end-of-unit (EOU) assessments administered to samples of 5th and 8th graders. We organized our findings around the three research questions that animated the general design of the pilot study. Our goal throughout was to collect information related to each of the guiding research questions to support, ultimately, revisions to the prototypes and to learn more about how three-dimensional end-of-unit tasks could be used in practice by teachers.

Our first research question focused on the utility of the EOU assessments for providing evidence of students' three-dimensional science learning. We collected information related to whether it was appropriate to use the EOU assessments for measuring students' science learning. What we found, briefly, is that while students were able to demonstrate science knowledge, there were some issues with the initial versions of the prototype assessments. Given that our plan was for each EOU to be administered in one class period, we discovered that substantial revisions to the tasks were needed because most tasks took students more than 20 minutes to complete, which meant, for the most part, students could complete only two of the three EOU tasks in a class period. While we expected to see some degree of missing responses from students, the number of missing responses by prompt (i.e., test item) was often much higher than we expected. Some of this may be because students simply ran out of time. We also found that several full classrooms skipped certain prompts or tasks within an EOU, suggesting that there were certain science topics that students were not familiar with or were not able to engage with on the assessment as intended.

Overall, the prototype EOUs were challenging for students in our study. While there were two assessments for which students were able to achieve the highest possible points, for most assessments, students fell short. The prototype EOUs did provide information about where students stood with respect to the rubrics scoring scheme used, and they also allowed us to measure variation in students' achievement as we

found prompts, tasks and EOU scores distributed across a range of performances. Importantly, based on the data obtained the Project has subsequently made adjustments to the timing and difficulty levels of the prototypes.

Further study will be needed to determine how well the end-of-unit assessments were able to reflect students' opportunities to learn. Throughout the pilot study teachers reported on whether they taught a particular topic, but there was no information on how deeply they went into a topic or how the topic was taught. While we found some evidence of differences in scores based on if teachers indicated they taught a given concept or not, these differences did not always favor the students who received instruction related to this concept. However, this could be due to differences in the organization of classrooms, or to the degree or depth to which the concept was taught.

Finally, while teachers were able to provide scores on student work, further study is needed to determine the reliability of these scores, particularly if the goal is to compare students across classrooms. While data on scores from different teachers on the same set of students were collected, these data were limited, and we saw differences in the overall reliability of scoring depending on the prompt or task being scored. While the limited pilot study data indicate we were able to see differences between and among students, and that some students were able to demonstrate their science knowledge, further information on how future iterations of the assessments will be used in the classroom need to be gathered to guide additional explorations into the design and use of the assessment tasks.

Our second research question asked if we could develop latent variable measurement models that fit the empirical EOU assessment data. Each of the prototype EOUs was scaled separately using the Rasch model, i.e., a one parameter IRT model. This modeling approach produced reasonable estimates of the items' difficulty parameters and student ability estimates. When using the Rasch model, item (or prompt) fit statistics were estimated which, in turn, proved useful for evaluating the measurement quality of the EOU prompts. Further, these fit statistics offered insights into the relationships among students' abilities and their responses to specific EOU prompts. More specifically, the fit statistics generated by the Rasch model measured the appropriateness of a prompt's difficulty relative to the students' abilities. Lower than expected values indicated that the prompt may have been too easy for our sample of students, leading to a high probability of correct

responses. Conversely, a higher-than-expected value suggested that the prompt may be too difficult. This model fit information was shared with the designers of the prototypes as they worked to improve the measurement quality for the next iteration of the EOU assessments.

The Rasch model fit statistics allowed us to evaluate the fit of a prompt or task in a more general sense, i.e., reflecting how well a prompt performs across the entire student ability spectrum. The use of latent variable models, like the Rasch model, allowed us to identify prompts that performed erratically suggesting that students' performance on the prompt may have been influenced by factors other than the students' abilities, such as guessing or simply misunderstanding the prompt. With this approach we were also able to flag prompts that were too predictable and, therefore, did not discriminate sufficiently among students with different abilities. In sum, our approach to latent variable modeling provided rich information about the measurement characteristics of the prototype EOUs. Unlike typical statewide assessment programs used for accountability purposes, IRT derived scale scores did not play a major role in this pilot, and thus were not computed based on a theta to scale score conversion formula.

Our third and final research question had to do with what the EOU assessment results tell us about students' science learning? As part of the investigation into this research question we examined the relationship between student scores and additional variables, including gender, prior ELA and Math learning, and curricular materials. We found that three out of the eight EOU assessments had statistically significant differences based on gender (in favor of females), but the sample size for this was low and so further study is needed to draw more solid conclusions. We also found that scores on the assessment tended to increase as prior ELA and mathematics levels increased. While this could indicate a dependency between ELA and math ability and the science assessment, there is often overlap between the science practices and ELA skills (e.g., communicating information) as well as the science practices and mathematical practices (e.g., problem solving). Therefore, more exploration is warranted to determine if there is too much of a dependency among and between skills.

Our analysis found statistically significant differences between students who used different curricular materials at the 5th grade (and for the Grade 8 EOU 2 assessment). However, without further investigation of the differences among the different curricula materials it was not clear how to interpret these differences. Further investigation to determine if the differences are due to desirable

characteristics (e.g., if different curricula cover different aspects on the assessment, we would naturally expect different scores) or to characteristics we would want to address in the assessment (e.g., if different curricula use different representations and the assessment is too closely aligned to one specific representation).

Cross-EOU Growth. The pilot study sample was modest—not all students in a grade took all four EOUs. Nonetheless, 64 5th graders and 21 8th students took all four EOUs. Based on these limited data we found that an increase in performance level from EOU to EOU reflected growth in students' learning because (a) each EOU had a unique set of performance level descriptors (PLDs) that form the basis for the task-PLD alignments and score estimations and (b) each level of each EOU's PLDs reflected a common expectation for student performance relative to the EOU's instructional unit. For example, PLD level 3 reflected the minimal performance expected of all students following each instructional unit. Thus, each level was qualitatively comparable across the four EOUs. In summary, the calibration of each level of the PLDs to a common goal relative to the instructional unit supports the measurement of cross-EOU growth. The current study had a limited number of cases from which to evaluate the efficacy of the proposed growth metric—change in performance level from EOU to EOU. It is recommended that the efficacy of this approach be further evaluated when a more robust data set is available.

Reporting of the EOU results. In the case of the pilot study, teachers scored their own students, and thus had access to student level data. However, no additional data were reported back to teachers about their students, and additional guidance on how this information could be used to inform subsequent units of instruction were not provided. Nevertheless, the pilot results suggest EOUs scores could be used to report back to teachers. We explored whether two different reporting metrics might be used to summarize individual student performance for each EOU and aggregated across EOUs.

Students could receive a reportable performance level based on each administered EOU. These performance levels, for example, may be used for reporting individual student results from multiple EOUs. Profiles can be summarized at the individual student level by reporting performance level profiles in both tabular and graphical formats. Performance level results can also be reported at the group level for each EOU. Group level performance level results are typically reported as the percentage of students in the group attaining each level. Multiple EOU administrations can be reported at the group level by reporting the percentage of students in the group

achieving each level on each EOU in both tabular and graphical (e.g., stacked bar chart) formats. Performance level reports for multiple EOU administrations over the course of the year can be supported via Performance Level Profiles. For example, a rubric may be adopted that links students' four EOU performance level profiles with an overall performance level.

It is important to note that we did not have a common scale across EOUs in a grade. However, performance-level based scores can be reported for each EOU and aggregated across EOUs to support within-grade, cross-EOU score interpretation based on the following rationale: Each EOU has a unique set of PLDs that form the basis for the Task-PLD alignments and cut score estimation and each EOU's PLD level reflects a common expectation for student performance relative to the EOU's instructional unit. PLD-based scores can be averaged on individual student reports to summarize multiple EOU administrations. Group level scores can be reported as an average of the individual students' PLD-based scores.

Educators may use the PLDs to inform subsequent units of instruction. That is, educators are able to review the descriptor for a student's current level of performance on an EOU—this tends to describe the range of performance for students achieving that level. However, by examining the next higher level, the educator can observe the skills the student needs to acquire to advance to that higher level. While the subsequent unit of instruction may be quite different, the information obtained from such a review may provide insight into students' strengths and weaknesses to inform the next unit of instruction—see below for a brief description of the subsequently funded CASCIA Project's interpretive resources that were developed for each revised EOU.

D. Summary of SIPS' Accomplishments

SIPS was an ambitious project in pursuit of multiple goals, primary among them is integration of science curriculum, instruction, and assessment resources for multiple instructional units at each of two grade levels. Among its accomplishments was the integration of two major conceptual and principled design frameworks–Understanding by Design and Evidence Centered Design–to guide the creation of Curriculum-Instruction-Assessment Unit materials and Design & Development Tools together with a multitude of specific resources for each C-I-A element of eight science learning units. Despite its limitations, the Pilot study data collection was sufficient for determining the quality and variability

of student performance on challenging, multi-dimensional science assessment tasks. The data collection also proved sufficient for providing evidence regarding: (a) teacher capabilities for reliable task administration and scoring, (b) challenges students face in task completion time and comprehension, (c) guidance for EOU task revision and scoring for subsequent use and validation, (d) EOU basic measurement properties, (e) exploration of alignment of performance with claims associated with embedded standard-setting processes, and (f) suggesting ways to evaluate year-long performance.

Since the completion of SIPS, a follow-on project called CASCIA, also funded by the U.S. Department of Education and involving some of the original SIPS partners, has pursued EOU assessment revision based on the SIPS pilot study results together with the development of interpretive guides and resources for each of the revised EOUs. It is beyond the present chapter to describe the work being done in the CASCIA project to validate the EOUs and interpretive resources, as well as what they are learning about classroom implementation of the instructional units and EOUs. However, it is useful for present purposes to provide an illustration of the types of interpretive resources that have been created to support multiple stakeholders for understanding and using results from the EOUs. Figure 10 is an illustration of the types of interpretive resources CASCIA has designed and is making available, who they are directed towards, and their intended interpretive use. Further information about these resources and other findings regarding their use should be directed to members of the CASCIA Project team via edCount LLC.

Figure 10.
Examples of the Reporting Mechanisms Developed by the CASCIA Project.

Reporting Mechanism	Audience	Proposed Purpose / Uses
Individual Score Report (ISR)	• Students • Parents/Guardians • Educators	• Summarize individual student performance on the end-of-unit assessment that can be used to monitor student progress and pl meaningful learning opportunities to ensure students are on trac to achieve end-of-year learning goals in science.
Classroom Roster Report (CRR)	• Educators • Administrators	• Summarize student performance by classroom on the end-of-uni assessment and offer information about students' instructional needs levels that educators can use to inform a variety of individualized, small, and whole group learning opportunities an make timely and meaningful adjustments to instruction.
Interpretive Guidance and Instructional Strategies	• Educators	• Provide information to help educators understand their students performance on the end-of-unit assessment and offer instructio strategies and resources for planning and adjusting instruction to help students learn.
Family Guidance and Learning Resources	• Parents/Guardians • Students	• Provide information to help families understand their student's performance on the end-of-unit assessment and offer resources recommendations for engaging their student in science learning home.
Task Interpretation Guide	• Educators	• Provide information to help educators understand the assessmer tasks and prompts, their features, and the evidence they are designed to elicit about student learning, and to reflect on prior

IV. Lessons Learned and Implications for Future Work on Assessments to Support Teaching and Learning in Science

We began this chapter with a description of the changes in expectations for student knowledge and learning in science as signaled by the 2012 NRC *Framework for K–12 Science Education* report and the derivative 2013 *Next Generation Science Standards*. In addition to describing multiple dimensions of knowledge–Disciplinary Core Ideas, Crosscutting Concepts, and Science and Engineering Practices–these reference documents specified ways of knowing in the form of multi-dimensional performance expectations requiring their integration. The goal is to have knowledge capable of explaining scientific phenomena, solving problems, and designing solutions to challenges posed by the natural and designed world in which we live. The ensuing decade has seen multiple efforts to articulate the instructional and assessment challenges posed by this contemporary framing of science proficiency. The two projects we have overviewed in this chapter represent some of the many attempts to address these challenges with a particular focus on assessment design, implementation, and interpretation for students in grades K–8. What follows are some reflections on what has been learned and issues that remain to be addressed by the science education research, development and practice communities.

A. Challenges of Multidimensional Science Assessment Design

Early on, the challenges of multi-dimensional science assessment design were duly noted, and recommendations were made that developing valid and reliable assessments for formative or summative use in classrooms and for large-scale assessment at state levels would require application of a principled approach to assessment design. The NGSA and SIPS projects are illustrations of the benefits that accrue from following such advice, emphasizing application of the Evidence-Centered Design framework articulated by Mislevy and his colleagues. The assessments designed within each project have well specified claims as to what knowledge and skills are being assessed and what evidence is required in student responses to support proficiency. The design patterns and item specifications are transparent allowing for the tasks to reviewed by experts as to their validity and the interpretability of student performance. By following a principled design process, the stages of which have been articulated in both projects for their respective tasks, others can use these design tools to develop new tasks aligned to multiple aspects of the Framework or NGSS for various grade levels and content areas.

B. Challenges of Interpreting and Scoring Multidimensional Science Performance

One thing that we have not focused on in our discussion of the assessments developed under each project is the issue of how best to interpret performance on the types of multi-dimensional tasks developed by each project. Given that the tasks and performances are supposed to be multi-dimensional, many educators and assessment designers advocated for the production of "separate" scores for each of the dimensions represented in the task. For example, a score for the disciplinary content and a score for the science and engineering practice. We, however, have viewed such an approach as inappropriate and antithetical to the presumption of integrated knowledge that is useable. Thus, in both projects, the interpretation of student performance focuses on evidence of integrated proficiencies that vary in their sophistication relative to the target proficiency for the given task. This avoids sending a message to educators that instruction should focus on the dimensions as separable targets and maintains an instructional focus on dimensional integration during instruction. Based on our experience with teachers using our tasks, we continue to believe that this approach to interpretation and scoring is far more meaningful and useful for both formative and summative interpretive uses.

C. Challenges of Integrating Curriculum, Instruction and Assessment in Science

What educators need to advance their own instructional practice and their students learning in the ways demanded by the Framework and NGSS are coherent and integrated curricular, instructional and assessment materials and resources. Unfortunately, the vast array of science education resources available to teachers since the appearance of the Framework and NGSS are curricular materials with weak and inadequate assessment materials for formative and/or summative classroom use. The development of assessments for most curricular products is largely an afterthought with little to no attention to assessment development using a principled approach such as ECD. One of the major contributions of the NGSA and SIPS projects is bringing curriculum, instruction and assessment together to achieve greater coherence in the classroom. In the NGSA project this has come about by working with teachers to integrate the various tasks into their curriculum and instructional unit materials by providing explicit guidance as to what is being assessed and where it fits with respect to a progression of learning anchored against the NGSS performance expectations. The SIPS project has directly taken on the coherence and integration challenge by bringing together the Understanding by Design curriculum and instruction design framework with the Evidence-Centered assessment design framework. Thus, while SIPS does not claim to provide a complete curriculum, instruction and assessment "package"–a so called "shrink wrapped" solution–it does provide a wealth of resources that teachers can adapt to their contexts and needs as well as tools and examples for how this can be done for other units of instruction at varying grade levels. We cannot underscore the degree of challenge that the SIPS project encountered in bringing these design frameworks together and the benefits that have accrued in terms of the materials and models that have resulted.

D. Benefits of the Work

No project in the science education field can begin to address all the issues related to the teaching, learning and assessment of science proficiency as it has now been envisioned. Each of the two projects described here have limitations with respect to scope of the problems addressed and degree of contribution. Nevertheless, we offer the perspective that much has been accomplished for multiple audiences and stakeholders.

For the Assessment Design and Development field writ large, and in science education specifically, much as been learned about how to conceptualize and execute the design process for multi-dimensional tasks. Models have been developed in both projects that can be deployed by others and modified as needed to create new tasks whether they be multi-dimensional tasks requiring the integration of mathematics practices and content as required by contemporary mathematics standards, or for additional tasks and task types for science education use, including those that can be used on large scale state assessment and/or for classroom or state performance assessments.

For Educators, including State education and assessment leadership teams, District C-I-A leadership teams, and Classroom teachers, both projects provide specific resources that are ready for deployment as well as models and practice guides to support professional learning and additional resource development. We know that the NGSA resources are being used by thousands of teachers as part of their classroom practice and many are using the design guidance to develop new tasks and interpretive tools. We also know that educators in multiple states, including the lead state of Nebraska, are using the SIPS resources for ongoing instruction and as professional development resources with multiple districts.

Finally, one of the most important benefits of the work of both projects is for Students in our K–12 classrooms across the country. Students (and their teachers) now have challenging tasks that can help them develop an understanding and appreciation of what is expected of them with respect to science proficiency. When our assessments are used wisely with constructive feedback from their teachers, students can gain proficiency and confidence in their science learning. Hopefully, they can come to appreciate more fully the elegance of science as a disciplinary activity that goes beyond memorization of facts and procedures and see it as a way to understand their world and guide their personal decision making in many facets of life.

E. What's Needed and What's Next?

We have alluded to some of the many things needed in the field of science assessment and for these two projects. Perhaps the best way to sum up and consider what's next is with respect to concerns regarding validity. Any science assessment effort, whether it be the NGSA tasks designed for classroom formative use, or the SIPS EOU assessments designed for classroom and potential large-scale state use, a primary concern is evidence regarding the intended interpretive use of the resources.

While each project obtained various forms of evidence related to their validity arguments, much remains to be done. The evidence needed is of multiple forms and goes well beyond traditional quantitative measurement or psychometric results (e.g., Pellegrino, DiBello, & Goldman, 2016). While the latter are needed as part of the validation argument, far more of the desired evidence will come from the world of practice. In particular, we need to know far more about how and how well educators can use the NGSA and SIPS resources to impact their practice and consequentially the learning of their students. We are hopeful that future projects making use of the NGSA and SIPS resources will provide many aspects of that evidence.

References

Alozie, N., Haugabook Pennock, P., Madden, K., Zaidi, S., Harris, C. J., & Krajcik, J. S. (2018, March). *Designing and developing NGSS-aligned formative assessment tasks to promote equity.* [Paper presentation]. Annual conference of the National Association for Research in Science Teaching, Atlanta, GA.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement.* (CPRE Research Report Series No. RR-43). Consortium for Policy Research in Education, University of Pennsylvania.

Davies, S. (2010). *Effective assessment in a digital age: A guide to technology-enhanced assessment and feedback.* JISC Innovation Group.

Fuhrman, S. H., Resnick, L., & Shepard, L. (2009). Standards aren't enough. *Education Week, 29*(7), 28–29.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education, 101*(5), 854–867.

Gane, B. D., McElhaney, K. W., Zaidi, S. Z., & Pellegrino, J. W. (2018, March). *Analysis of student and item performance on three-dimensional constructed response assessment tasks*. Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multi-dimensional learning. *European Journal of Education, 53*(2), 176–187.

Gane, B. D., Zaidi, S. Z., McElhaney, K. W., & Pellegrino, J. W. (2019, April). *Design and validation of instructionally-supportive assessment: Examining student performance on knowledge-in-use assessment tasks.* Paper presented at AERA Annual Meeting, Toronto, ON, Canada.

Gorin, J. S., & Mislevy, R. J. (2013, September). Inherent measurement challenges in the next generation science standards for both formative and summative assessment. In *Invitational research symposium on science assessment*. Educational Testing Service.

Harris, C. J., Krajcik, J. S., & Pellegrino, J. W. (2024). *Creating and using instructionally supportive assessments in NGSS classrooms*. National Science Teachers Association.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education, 98*(4),674–704.

Lewis, D., & Cook, R. (2020). Embedded standard-setting: Aligning standard-Setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39*(1), 8–21.

Lewis, D. M., Mitzel, H. C., Mercado, R., & Schulz, M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*, (2nd ed.). Lawrence Erlbaum.

McElhaney, K., Vaishampayan, G., D. Angelo, C., Harris, C. J., Pellegrino, J. W., & Krajcik, J. (2016, June). Using learning performances to design science assessments that measure knowledge-in-use. In C. K. Looi, J. L. Polman, U. Cress, & P. Reiman (Eds.), *Transforming learning, empowering learners: Proceedings of the 12th international conference of the learning sciences* (ICLS) 2016, Vol. 2 (pp. 1211–1212). International Society of the Learning Sciences.

McElhaney, K. W., Zaidi, S., Gane, B. D., Alozie, N., & Harris, C.J. (2018, March). *Designing NGSS-aligned assessment tasks and rubrics to support classroom-based formative assessment*. Paper presented at the NARST Annual International Conference, Atlanta, GA.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D.W., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns*. Springer.

National Research Council (1996). *National science education standards.* National Academies Press.

National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Science Teachers Association (2016). *NSTA position statement: The Next Generation Science Standards.* https://www.nsta.org/about/positions

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

Pellegrino, J. W. (2010). The design of an assessment system for the Race to the Top: A learning sciences perspective on issues of growth and measurement. In P. Forgione & N. Doorey (Eds.), *Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda*. Center for K–12 Assessment & Performance Management, Educational Testing Service.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, 340, 320–323.

Pellegrino, J. W. (2018). Assessment of and for learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.). *International handbook of the learning sciences (pp. 410–421).* Routledge-Taylor & Francis.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education 43*(2), 119–34.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2014). *Developing assessments for the next generation science standards*. National Academies Press.

Pennock, P. H., & Severance, S. (2018, March). *Comparative analysis of three-dimensional research-based and classroom-based rubrics for formative assessment.* Paper presentation at NARST Annual International Conference, Atlanta, GA.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching,* (5th ed., 787–850). American Educational Research Association.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84.

Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom assessment system to support ambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.)*, Reimagining balanced assessment systems (pp. 93–131).* National Academy of Education.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education.* Boston, MA: Kluwer Academic.

Sezen-Barrie, A., & Kelly, G. J. (2017). From the teacher's eyes: Facilitating teachers' noticing on informal formative assessments (IFAs) and exploring the challenges to effective implementation. *International Journal of Science Education, 39*(2), 181–212.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21–34.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Zaidi, S.Z., Ko, M., Gane, B.D., Madden, K., Gaur, D., & Pellegrino. J. W. (2018, March). *Portraits of teachers using three-dimensional assessment tasks to inform instruction*. Paper presented at the NARST Annual International Conference, Atlanta, GA.

Zhai, X., & Wiebe, E. (2023). Technology-based innovative assessment. In C. Harris, E. Wiebe, S. Grover, and J. W. Pellegrino (Eds.), *Classroom-based STEM assessment: Contemporary issues and perspectives, (pp. 99–126).* Community for Advancing Discovery Research in Education (CADRE). Boston: Education Development Center.

Notes

The work described in this chapter spans a period of time from 2014–2023, with funding from multiple organizations including the National Science Foundation, the Moore Foundation, the Chan Zuckerberg Initiative, and the U.S. Department of Education. It is the product of many individuals representing multiple organizations.

The following individuals made significant contributions to the NGSA Project work: Joseph Krajcik, Christopher Harris, Daniel Damelin, Brian Gane, Sania Zaidi, Diksha Gaur, Kevin McElhaney, Nonye Alozie, Phyllis Pennock, Sam Severance, Krystal Madden, Angela DeBarger, Carla Strickland, Debbie Leslie, Jeanne DiDomenico, Diksha Gaur, Samuel Arnold, and Elizabeth Lehman.

The following individuals made significant contributions to the SIPS Project work: Ellen Forte, Erin Buchanan, Bill Herrera, Charlene Turner, Rhonda True, Daisy Rutstein, Allison Kaczmarski, Donald Wink, Brian Gane, Sania Zaidi, Mon Lin Ko, Mary Nyaema, Daniel Lewis, and Nathan Dadey.

Formative Assessment in a Digital Learning Platform

Kristen DiCerbo

Khan Academy

Abstract

Students engage in practice on digital learning platforms and are able to receive both necessary scaffolding to build their skills and immediate feedback to course correct. In addition, these platforms are able to collect and aggregate information from these practice experiences, becoming formative assessment tools.

Platforms like Khan Academy collect and analyze student performance data—such as accuracy, scaffold usage, and time spent—to provide skill-level insights that inform instructional decisions. Unlike traditional standardized tests, digital assessments prioritize real-time feedback and continuous learning over one-time evaluations.

These systems support motivation by encouraging students to persist through practice, reinforced by features like streaks, progress tracking, and visible mastery indicators. Additionally, real-time feedback mechanisms help students understand their current proficiency and determine their next steps for improvement. Teachers also receive actionable insights, although challenges remain in integrating this data effectively into instruction.

The chapter further explores the potential of generative AI, such as Khanmigo, to enhance assessment experiences, including the skills we assess, how we assess them, and how users understand the results of the assessment. However, ensuring data reliability and meaningful feedback remains an ongoing challenge, emphasizing the need for continued research in AI-assisted assessment.

Formative Assessment on a Digital Learning Platform

We live in a world where students engage in practice on digital learning platforms and are able to receive both necessary scaffolding to build their skills and immediate feedback to course correct. In addition, these platforms are able to collect and store information about these practice experiences, both the correctness of student responses and information about scaffolds used, attempts taken, and time spent. In distilling this information, digital learning platforms such as Khan Academy provide meaningful summaries about what students know and can do at a level of granularity that can help inform instructional decisions, essentially becoming formative assessment[1] tools.

Time Versus Information

Schools, districts, and states want to cut testing time as much as possible, and they pass this pressure on to the organizations that develop assessments. Computer adaptive testing (CAT; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000) emerged as one way to reduce the amount of time students spend on a single assessment. Instead of asking every student a set number of questions, the CAT selects questions to maximize the information it learned from their responses and can produce an overall math achievement score with fewer questions.

However, an overall math achievement score, or broad subdomain scores, is not particularly helpful in making day-to-day instructional decisions. Even information at the standards level is often deemed too coarse for deciding which students should do what in a given lesson. To be instructionally informative, teachers need information at a skill level. There is really no way to ask enough questions to reliably measure individual skills on a single assessment without that assessment lasting an inordinate amount of time.

Traditionally, teachers have filled the gap between the information they get from standardized assessments and the granularity of information they need to make decisions by creating and administering their own assessments. These can be anything from exit tickets (1–2 questions asked at the end of a class period that the teacher collects as students leave class and reviews to determine if the main

1 Formative assessment is the range of formal and informal procedures used in classrooms to help teachers and students understand learning while it is in process in order to adjust teaching and learning strategies. It stands in contrast to summative assessment, which is conducted at the end of a segment of learning in order to understand whether a learner has achieved the intended outcomes.

point of the day's lesson was achieved) to quizzes and unit tests. While these short teacher-made assessments serve to give teachers more information, they also have downsides. First, they may not be aligned to the state standards and assessments, leading to the feeling among students that there is a mismatch between what they are learning and what is ultimately assessed. Second, constructing good quizzes is time-consuming and having every teacher create their own is yet another burden given to already overworked teachers. Third, teachers have to score the assessments, and little information about the students' performances is captured for either longitudinal tracking or communication with anyone outside the classroom (for example, the next year's teacher).

The rise of digital learning platforms offers another alternative for gathering skill-level information about what students know and can do. Students are engaged in skill-level practice on platforms such as ALEKS, i-Ready, IXL, and Khan Academy, often for 60+ minutes per week. Information about their performance, including their responses to individual activities, scaffold use, and feedback is all captured and stored. Responses are automatically scored, and student-level, class-level, school-level, and district-level information about performance is available in real-time. These platforms are thought of as instruction, practice, and learning platforms. However, they are best understood as learning and assessment platforms. Digital platforms that offer students the ability to answer questions, capture and aggregate information from those performances, and use that information to make recommendations about future instruction and learning are functioning as formative assessments.

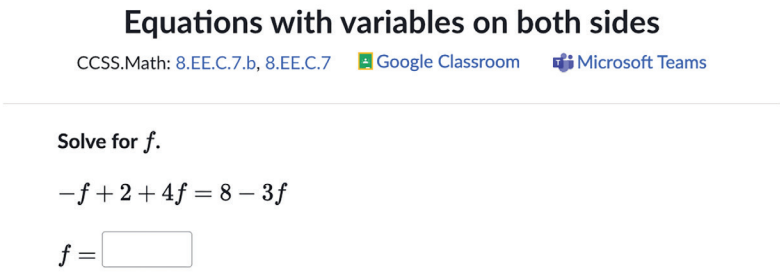
A Brief Overview of Assessment at Khan Academy

To provide context for the following discussion, here is a brief overview of the learning and assessment system at Khan Academy. All learning and assessment experiences draw from a bank of 120,000-plus activities, mostly traditional items (as shown in Figure 1) but also some projects, particularly in computer science. The following are the types of experiences on which students solve problems and get feedback:

- Exercises consist of 4–11 items/activities, all focused on a single skill.
- Quizzes and unit tests cover multiple skills (2–4 for quizzes and 5–10 for unit tests).

- Course challenges cover content from the entire course. They are sometimes used at the beginning of courses as diagnostic tools but more commonly at the end of courses. They are also frequently used as preparation for other end-of-year summative exams.
- Mastery challenges are a means of engaging learners in spaced practice. They sample questions from skills the learner has already mastered.

Figure 1.



The activities for each experience are drawn randomly from the pool of all activities aligned to that skill. The random draw gives the system a lot of technical simplicity; there is no need for in-production item selection based on the statistical properties of the item or on-the-fly statistical computation of learner skill. The downside is that some learners may get a series of easier or more difficult questions by chance, which is addressed by: 1) setting the high bar for reaching proficient status (100% of items correct on an exercise), 2) allowing as many attempts as students wish on an exercise, quiz, or test, and 3) writing questions to be of similar difficulty level and monitoring item statistics.

Mastery Learning

Mastery learning is an approach to instruction that emphasizes students engaging in instruction and practice until they reach the defined level of proficiency (See Guskey, 2022, for a comprehensive overview). It is commonly defined as a cycle where students: 1) are assessed to determine what skills they have and have not

mastered, 2) engage in learning activities on skills they have not mastered, and 3) are re-assessed on those skills. The instruction-assessment loop continues until mastery is achieved. At Khan Academy, Mastery Learning means ensuring that learners have the opportunity and incentive to master the skills they need to prepare them for future learning. Learners continue to work on a skill until they reach a given level of proficiency or performance. In a mastery learning system, no assessment is meant to be "your final chance to demonstrate your knowledge." There are no limits on how many attempts learners get on exercises, quizzes, or course challenges.

Setting Expectations for Progression

Expectations for progression are built into the foundation of Khan Academy's mastery learning system, which defines a series of levels, from "attempted" to "familiar" to "proficient" to "mastered." Learners advance through these levels as they get more questions correct on exercises, quizzes, unit tests, and course challenges.[2] Skill mastery rolls up into unit mastery and course mastery. Teachers can assign unit and course mastery goals for students. For example, if the class is working on negative numbers for the next three weeks, a teacher can create an assignment that challenges the students to get to proficient or mastered status on the 16 skills related to negative numbers (e.g., negative numbers on a number line, ordering negative numbers, etc.) by the end of week three.

When deciding how to define mastery, we had the options of 1) using underlying probabilistic models of mastery and defining cut points for each level or 2) creating human understandable rules for progression. We settled on creating rules for progression. For example, to get to proficient status, students can either get 100% of questions right on an exercise or, if already at familiar status, get questions on that skill correct on a quiz or unit test. There were two factors in the decision to use a rule-based, rather than probabilistic system: user preference and having a meaningful signal from the score. Students were clear: they wanted to know what they had to do to achieve mastery at each level. When working with an underlying probabilistic model, students have to keep working until the model tells them they have reached a level, but they do not know if they need to do 5 more problems or 10 more problems until they hit a level. They keep answering questions without

2 Learn more about Khan's mastery mechanics here: https://support.khanacademy.org/hc/en-us/articles/115002552631-What-are-Course-and-Unit-Mastery

understanding how that impacts their progress toward mastery and report significant frustration with what they perceive as a black box.

The question then becomes whether the rule-based system provides a good signal of mastery. Khan Academy has an offering called MAP Accelerator where: 1) students take NWEA's interim MAP Growth assessment in the fall, winter, and spring, 2) their score feeds into Khan Academy, and 3) they get placed in content at their level. The sharing of score data between the systems means that we are able to match students practicing on Khan Academy with their NWEA growth scores over a year. Analyzing the data revealed a significant relationship between the number of skills on which students get to proficient status and their increase in MAP Growth scores (Yamkovenko, 2023). Similarly, a third-party study (Oreopoulos, Gibbs, Jensen & Price, 2024) showed that learners in Texas who leveled up an average of 3+ skills per week showed significant growth (effect size = .24) on the Texas STAAR test and that the relationship between skills per week and STAAR growth continued linearly.

The mastery system also allows the investigation of whether learners should work on more skills, getting them to familiar status, or fewer skills but getting to proficient status. As previous research on mastery learning would suggest, getting to the higher level of proficiency, even on fewer skills led to greater gains on the MAP Growth assessment than getting to the lower level of familiar on more skills (Yamkovenko, 2023). One of the keys to a mastery learning-based system is to set a high standard for what it means to get to mastery. Previous meta-analyses have suggested "the higher the better," with mastery scores of 100% showing better retention over time than mastery scores set at 80% (Kulik, Kulik, & Bangert-Drowns, 1990). Our findings, consistent with what we would expect from theory and previous research, gave us confidence that our system of progression based on understandable rules, does provide a clear signal about student achievement and progress.

Supporting Motivation and Engagement

Learning is hard. Applying the attentional resources and cognitive effort required to engage with new material and to continue practicing until mastery levels are reached challenges students. Like many learning and assessment experiences, Khan Academy is challenged to motivate and engage learners.

We know from basic motivation research that mastery goals lead to better motivation than performance goals (although not always higher achievement; Senko, 2019). Mastery goals focus on improving one's own performance relative to intrapersonal or absolute standards, while performance goals focus on outperforming interpersonal or normative standards (e.g., getting the highest score in the class). The idea of getting more skills to proficient aligns well with the existing research as proficiency is a standard and reaching it for a number of skills is an intrapersonal goal. As such, we have used it as a basis for a number of motivation mechanics. First, we have a "streaks" system which tracks the number of weeks in a row that a student levels up at least one new skill to proficient and encourages students to keep their streak going. Second, we have a levels system where students move up levels as they get more skills to proficient. Finally, we have visual representations of learners' mastery status on all skills in a course. At the top of each course page, there is a graphic that provides a representation of each skill in the course that gradually fills in as students move from familiar to proficient to mastered. There was a significant increase in student practice activity following the introduction of the visual tracking feature.

Feedback

The key purpose of formative assessment is informing instructional and learning decisions. For students, formative assessment helps them decide what to work on next, for example whether to keep practicing a skill or move on to the next. For teachers, it means providing insight both on what to assign on the platform and what to do in the classroom. Research has shown mixed results for the impact of feedback on learning. Meta-analyses of the impact of feedback on learning report overall positive results, but significant heterogeneity across studies (Wisniewski, Zierer, & Hattie, 2020). A closer read reveals that the nuances of how feedback is delivered, when, and the content of the message all influence the effectiveness of feedback (Shute, 2008).

Students engaging in learning and assessment on Khan Academy receive feedback after completing each item. On multiple choice questions, if a student selects an incorrect option, they are told it is incorrect and given a 1–2 sentence rationale for why the option is incorrect (See Figure 2). It is important that these explanations are short and easily understood, presented in what Shute calls "manageable units" (Shute, 2008, p. 177). The student is given the option of trying again. If the

student answers correctly, they get an indication that the response is correct. For numerical response items, the student is given correctness feedback (i.e., correct or incorrect). Students who get the item wrong are given the option to either retry or view the worked solution (See Figure 3) and move on. Originally, viewing the worked solution was optional but we now show it to everyone who selects moving to the next question because we want all students who are not trying again to see it. The feedback immediately follows the student's response on a specific question so that it can influence their understanding and behavior on the next question on that skill. Most assessment experiences do not provide immediate item-level feedback, in part due to its potential impact on motivation and learning. If the ultimate goal is to measure students' understanding at a point in time, instantaneous feedback could change the students' understanding and thus interfere with the measurement. In the Khan Academy experience, the primary goal is learning. Due to the mastery learning mechanics, students understand that they have multiple chances to show what they know. The focus is put on mastering the skill, not getting a particular score on their one chance to take a test. As a result, the potential demotivational impact of receiving feedback that they are not correct is softened and we hope it does change their understanding of the topic.

Figure 2.

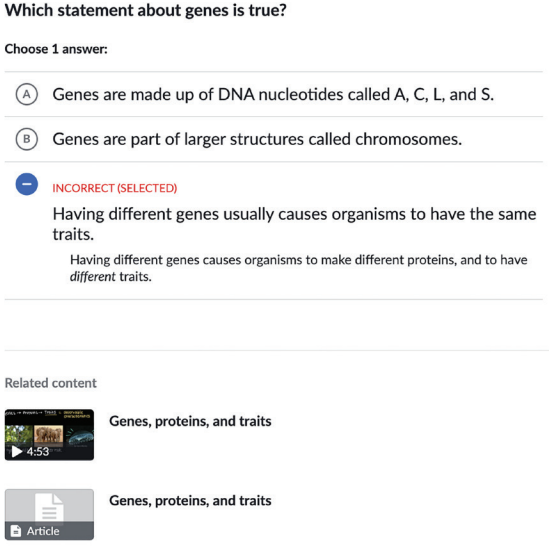


Figure 3.

Solve for m.

$-7 + 4m + 10 = 15 - 2m$

$m =$ ☐

1 / 3 We need to manipulate the equation to get m by itself.

2 / 3 $-7 + 4m + 10 = 15 - 2m$

$4m + 3 = 15 - 2m$	Combine like terms.
$3 + 4m + 2m = 15 - 2m + 2m$	Add $2m$ to each side.
$6m + 3 = 15$	Combine like terms.
$6m + 3 - 3 = 15 - 3$	Subtract 3 from each sid
$6m = 12$	Combine like terms.
$\frac{6m}{6} = \frac{12}{6}$	Divide each side by 6.
$m = 2$	Simplify.

3 / 3 **The answer:**

$m = 2$ Let's check our work!

Once the student has completed an exercise, quiz, unit test, or course challenge, they are immediately given a performance summary in an easy-to-understand indication of how many questions they got right and the total number of questions. They are then told the skills on which they changed mastery status. Based on this information, students are able to choose whether they would like to revisit

instruction and practice on skills that they have practiced but not reached mastery on or proceed to the next skills in the progression. The important point is to provide students with an understanding of their current status on each skill so they can make informed decisions.

Similarly, teachers are given multiple ways to view student results. There is a traditional "score report" on which teachers can see both the most recent and best scores students have gotten on exercises, quizzes, unit tests, and course challenges. Although Khan Academy focuses on skill mastery, the majority of schools still have a system that requires the reporting of average scores. Therefore, the experience offers traditional score reporting as an option for practical reasons. In addition, teachers can use a more mastery-based approach. They are able to look at a skill view, which shows the mastery status of all the students in their class on particular skills (See Figure 4 and Figure 5). They can look at the mastery status of individual students across a group of skills (See Figure 6). They can also get summary information on the number of skills students have leveled up on in a given time period. Teachers are also able to get item-level reports that summarize the performance of the class on individual items, including the percentage of students that selected incorrect answers on multiple choice questions.

Figure 4.

Figure 5.

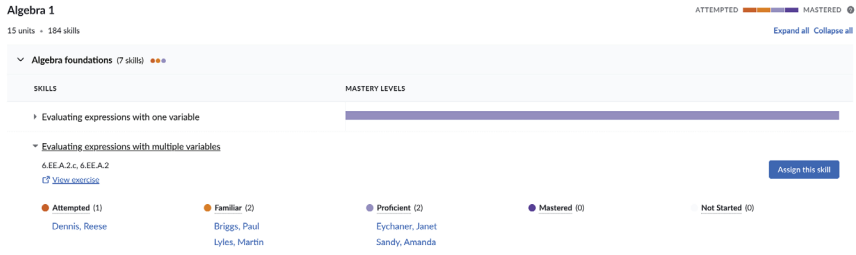
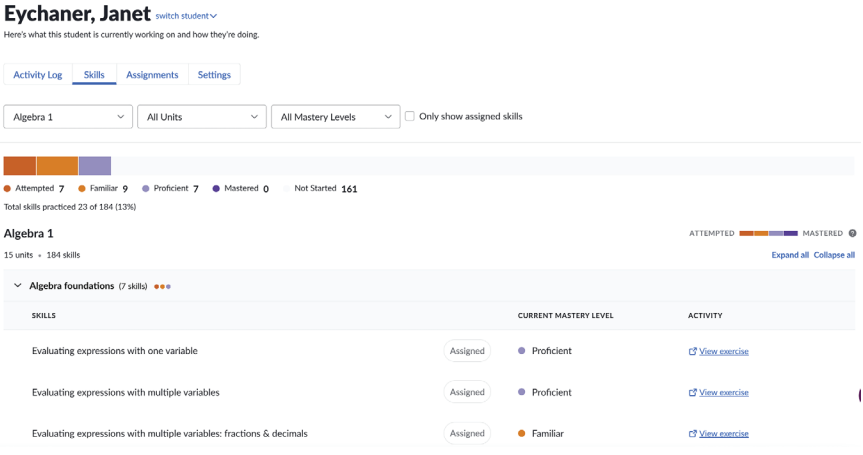


Figure 6.



At no time does the reporting at Khan Academy offer a norm-referenced score or any kind of scaled score. Scaled scores are scores that adjust a student's raw score for the difficulty of the exam they took. The purpose of reporting at Khan Academy is to communicate progress and clearly suggest actionable next steps, and we have not found scaled scores to be helpful in reaching those goals. While we sacrifice some precision, research has suggested that simple percentage-type scoring correlates well with other methods, including Item Response Theory and Bayesian network-based scoring (Choi, DiCerbo, Ventura, Lai, Wood, & Iverson, 2019). We are willing to sacrifice the precision given our

low stakes environment in order to gain simplicity (including technical simplicity in building the platform) and interpretability.

This section on reporting would not be telling the whole story without acknowledging the relatively low numbers of teachers who access reports. While educational technology companies do not tend to publish these numbers, completely unscientific, informal questioning suggests that likely fewer than 30% of teachers regularly look at student data on digital learning platforms. Teachers are notoriously busy, plus they often interact with a variety of platforms that all report data individually and differently. Qualitative feedback from teachers suggests that they want not just the data reported but also recommended actions that they can either accept or reject. For all the talk about driving data-based decision-making, even with more granular, specific data, there is work to be done to inform instruction.

Does it Work?

The primary efforts to establish the efficacy of the use of Khan Academy center on demonstrating the learning gains that are associated with use. A collection of experimental, quasi-experimental, and correlational studies with statistical controls have consistently demonstrated that students who use Khan Academy at least 30 minutes per week and/or increase the number of skills on which they reach proficiency have better math performance on standardized tests, better grades, and increased college readiness (Grimaldi, 2023).

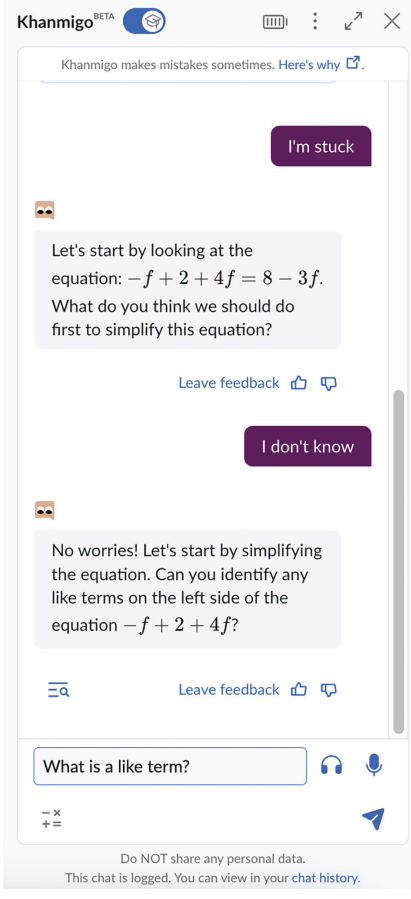
From an assessment perspective, the question of "does it work?" refers to validity, reliability, and fairness. Classroom formative assessments, especially teacher-made assessments, do not often undergo the close analysis of these factors to the level of rigor of summative assessments. The consequences of less psychometrically sound assessments are relatively small in the formative assessment space; a student might do a few extra problems practicing a skill, for example. There are two types of validity evidence that are of interest to users of formative assessment: evidence based on content and evidence based on external measures. School decision makers want to know if online practice systems align to state standards and whether they will predict performance on end of year assessments. In the case of Khan Academy, the course development process includes coverage maps to the academic standards being taught and assessed and the information is available by state on the Khan Academy website. There is

also evidence, as described above, that the levels of proficiency attained on Khan Academy math courses are significantly correlated to scores on other external measures of the same math achievement construct. The correlations hold across racial/ethnic groups and socioeconomic status. The close alignment between Khan Academy performance and summative test performance is not surprising, given the similarity in the format of items and exercises between the two. At least for now.

Generative AI and the Future of Assessment

In the fall of 2022, Khan Academy received a sneak preview of a large language model that we now know as GPT-4, and it significantly changed the direction of the Khan Academy offering. Large language models are a type of generative AI that produces text based on patterns it has learned from ingesting vast amounts of written content. The language generation means we can have conversational interactions with the AI in ways that have never been possible before. In March of 2023, Khan Academy released Khanmigo, an AI-powered tutor for students and assistant for teachers.

Khanmigo takes the power of the generative AI model and designs a specific education experience with it. For example, in math tutoring on Khan Academy, when a student wants to use the Khanmigo tutor, the student's input gets sent to the model along with instructions on how to act like a tutor. These instructions are based on research on what makes a good human tutor, which has been conducted over the past decades of trying to develop intelligent tutoring systems (e.g., Graesser, Person, Magliano, 1995). As a result Khanmigo, makes goal statements, course corrects when students are headed down the wrong path, and makes similar tutor moves that help the student get to the answer themselves (See Figure 7). In a writing coach application, separate instructions are sent to the model so it can evaluate various aspects of student writing (e.g., the ability of the introduction to capture attention, the use of evidence in an argument, etc.), provide that to students, and then engage in conversation about how to improve.



At the time this chapter is being written, there is significant excitement about the potential for generative AI to impact education, but it is early days in its development and it is not clear how much of the promise will be realized. In particular, education has typically moved slowly to adopt new innovations (Reich, 2020), often for good reason; the education of children is important and should not be subject to every fad that arrives on the scene. Ideally, evidence would be gathered on new interventions before they are scaled widely. Unfortunately, conducting the kinds of rigorous research done in, for example, the medical field

is also fraught with roadblocks and difficulty. As a result of numerous constraints, many decisions about which educational technology products to use are often based on word of mouth among district administrators and less information is gathered about effectiveness than would be desirable (Morrison, Ross, & Cheung, 2019).

In the current situation with generative AI, in the 2023–24 school year 53 school districts participated in a pilot of Khanmigo. Many did so with specific schools, domains, and/or grades in an effort to try out the tool before bringing it to scale. The state of Indiana released a request for proposals that allowed districts to receive funding for such pilots and also then ran teacher surveys to gauge their perceived usefulness (Appleton, 2024). The uses of Khanmigo clearly fell in the learning space, but give us some direction of how generative AI might impact the future of assessment.

Skills to Assess

Evidence-centered design (Mislevy, Almond, & Lukas, 2003) defines the domain model as the set of knowledge, skills, and attributes to be assessed. It is possible that the advent of generative AI opens up a new set of skills that should be assessed. In the workforce, many professionals are already using generative AI as an assistant. Software engineers at Khan Academy use the GitHub copilot to code with them. The engineer indicates what code they want to write, the AI copilot drafts code, and the engineer reviews and revises it. Similarly, many people who need to write text ranging from marketing emails to job descriptions are using the AI technology to create first drafts. The number and type of uses of generative AI suggests the importance of the skills of evaluation (of code and text) and editing is going to increase. However, evaluation and editing are rarely assessed currently, but should be considered in assessment research and development spaces.

New Task Models

The task model is the abstraction of the activity with which the person being assessed engages. Historically, the activity types used for assessment have been limited to what the technology available could support for large-scale automated assessment. When the only option available was optical scanners, multiple choice questions provided the best way to score a significant number of assessments quickly. As technology has progressed, variants, often called "technology enhanced items" appeared, including drag and drop, match and order, and more recently graphing, hot spot, and audio and voice items.

There has been a significant amount of work done on simulation and game-based assessment (e.g., Gobert & Sao Pedro, 2016; Shute & Ventura, 2013; Baker, Dickieson, Wulfeck, & O'Neil, 2017). Both offer the possibility of more authentic tasks for students. The premise of these assessments is that rather than ask students a question about how they would do something, we can ask them to do that thing in a simulated environment. Ideally, the use of tasks like those in the real world should shorten the assessment's inferential distance (Behrens, DiCerbo, & Foltz, 2019), the theoretical distance between what we observe someone do and what we infer about what they know and can do from that evidence. For example, there is a relatively large distance between observing which option was selected in a multiple choice question about computer networking and inferring that someone can configure a network. By offering students a simulation, we can observe them engage in many of the actual tasks we are interested in assessing (Behrens, Collison, & DeMark, 2008).

There are downsides to the use of more authentic types of assessment tasks. First, the assessment time-to-evidence ratio can be high. In games, it is often the case that students engage in 30–45 minutes of gameplay and only generate a few pieces of evidence that provide information about the construct of interest. SimCityEDU, a game-based assessment of systems thinking in which students worked to diagnose why city residents were sick and fix the problem, demonstrated the trade-offs between game play and evidence. Ultimately, the problem students needed to diagnose was air pollution; the solution was adding new energy types and removing some of the coal-burning power plants. In terms of systems thinking, evidence consisted of things like placing in new energy types before removing the coal plants vs. removing the coal plants first (which would leave the city with no power, an indication of poor systems thinking). However, it took a significant amount of game play to diagnose the problem causing the residents' illness and then to uncover solutions. After extensive gameplay, only a few pieces of evidence ended up in the statistical models estimating systems thinking proficiency (Castellano et al., 2014).

SimCityEDU also highlighted that inferential distance remains in many simulations and games. In analyzing moves students made in the game, it became apparent that about 5% of students were bulldozing the entire city. What should we infer from this behavior in regard to systems thinking? Were they thinking about rebuilding the city from the ground up based on ways to eliminate air pollution? Was bulldozing

the city the ultimate in systems thinking? As we found out when we asked a handful of students, it was mostly because bulldozing is fun. We find with simulations and games, in many cases, to eliminate inferential distance, we need to use language and ask students what they are thinking.

The need for new functionality is especially apparent when the skill to be assessed is rooted in language, such as collaborative problem solving. PISA (Program for International Student Assessment) undertook an assessment of collaborative problem solving skill in 2015, including producing a detailed framework describing the skill (Foster & Piacentini, 2023; OECD, 2017). The team wanted to observe students actually engaged in collaborative problem solving. Doing that with other human learners though was technically difficult and introduced a considerable amount of variability. So, the assessment had learners interact with automated agents. However, due to the difficulty of processing and scoring natural responses, students were given multiple choice options from which to choose a response, rather than entering a free-form response. If the students were able to type anything they wanted in a response, there was no good system by which their automated collaborators could engage in conversation about the wide range of things the students might say. Drafting these dialogue trees was a large task even in the agent-based solution. The assessment led to informative results. However, the difficulty in managing language continued to result in a gap between what was observed and the inference to be made.

Over the decades, significant work has been done on automated essay scoring (Shermis & Burstein, 2013). Work that began with the identification of features that correlated to human scoring, such as essay length, matured into models that used the meaning of words to evaluate essays. Today, many programs score essays at the same level of agreement to humans that other humans do, not perfectly because humans also disagree, but at a high level. The difficulty with these programs is that they usually require training the model on the specific essay to be scored using hundreds or thousands of human-scored examples. Additionally, from a learning perspective, just getting a score is not sufficient to help a learner know how to improve.

Enter generative AI. It cannot solve all of the problems, but it can, in combination with solutions we already have, improve our existing assessments. First, the models, with proper prompting can engage in conversation. Those skills involving

dialogue can potentially be assessed directly rather than through selected responses. Additionally, instead of trying to infer why a student did something a certain way, whether answering a traditional math problem or bulldozing cities, the AI can ask them and engage in dialogue about what is being observed. The inferential distance can be further closed.

Generative AI can also be good at giving feedback on writing. In classrooms currently, students do not often get assigned longer essays because they require a large effort to grade. As a result of that effort, feedback is often delivered many days or weeks after the writing was done and little feedback likely influences performance on future writing assignments. Khan Academy now has a writing coach feature that walks students through writing assignments and then provides feedback on aspects of the students' drafts. For persuasive essays, for example, Khanmigo gives feedback on students' introduction, use of evidence, structure, conclusion, and tone and style. The ability to provide feedback specific to elements of essay writing won't happen "out of the box" with a large language model, but it can with applications specifically designed to use the models for education. To get Khanmigo to provide this feedback, we split each area of feedback into a separate prompt. Each of those prompts contains instructions based on what writing teachers look for in that element, telling the model what to look for. Designers and engineers then created the means by which students could edit and Khanmigo could "see" the changes that students are making and converse with the students about them. The feedback functionality of generative AI has the potential to fill in the feedback gap in most automated scoring of writing.

Finally, generative AI has the potential to allow for more individualization of activities in ways that will enable for different background knowledge and experience to be considered. Even with the computer-adaptive test, the adaptivity focuses on the students' measured achievement level and the difficulty of the items. All students are working from the same item pool. However, we know that questions can be differentially difficult depending on familiarity with the (sometimes irrelevant) context of the question. The classic example of the impact of background knowledge on comprehension used for those familiar with the majority American culture is to give a reading passage about baseball and one about cricket. Americans with a deeper knowledge of baseball and nearly no knowledge of cricket do much worse on reading comprehension questions about cricket.

It is possible that generative AI could be used to adjust the background knowledge of questions for students in ways that do not penalize some students for their lack of familiarity with contexts. The adjustment of context in an assessment question has been impossible because it did not make sense to assume familiarity with given contexts based on someone's rough demographic profile. Now, it is possible that AI solutions could be used to tackle the problem of personalization. At Khan Academy, students can choose to converse with Khanmigo about their interests. Khanmigo probes on different topics, from food to sports to hobbies, and records up to 10 interests in the student's profile. Students can always go in and modify or delete what they have entered. These interests are then injected into different prompts to guide Khanmigo so the conversations can incorporate the interests. Currently, Khanmigo can adapt questions during a conversation to incorporate these interests. Still, the responses to the adapted question do not feed back into the mastery system, largely because we have not built the infrastructure for information from conversations to be incorporated into scoring and mastery mechanics. That said, there is a clear research and development need for mechanisms by which to equate items with differing contexts, potentially created in the moment of administration.

Reporting

As mentioned above, many teachers do not make use of data from digital systems. Despite valiant efforts at design research (Zapata-Rivera, 2018; Zenisky & Hambleton, 2015), score reports primarily do what their name suggests, and report scores. Generative AI offers the potential to let consumers of assessment results have conversations about the results, including asking questions about what they mean and getting recommendations from them. At Khan Academy we now have an AI tool for teachers called Class Snapshot where Khanmigo first gives a summary of student performance in the class, including the time spent and skills leveled up. The statistical summary is done with a calculator and fed to the large language model in order to ensure mathematical accuracy. The teacher can then ask questions such as “who needs help adding fractions?,” “who should I group together for a lesson on multiplying decimals?” and “what should I assign to my students next for practice?” The latter will produce groups of students of similar skills and suggest Khan Academy content. The teacher can then interrogate the model's responses and make decisions about whether to accept the recommendations, allowing teachers to obtain, not just data given to them, but clear options for action based on that data.

Challenges to Psychometrics

As the field starts looking at assessment in technology rich environments, some of the existing rules and procedures may need to be revisited (DiCerbo, Shute, & Kim, 2017). Many of the techniques used for measuring the psychometric properties of assessment were developed in the context of standardized assessment, consisting of discrete items specifically designed to assess a single construct, scored as correct or incorrect. Much of the evidence gathered from assessment in technology-rich environments (e.g., time spent, sequence of events) is not scored as correct/incorrect and often relates to multiple constructs. In addition, there is often variability in what activity is presented to different learners depending on their own progress and choices.

As described above, currently Khan Academy proficiency is a strong predictor of external assessment performance. As new methods of assessment are developed, the standards for acceptable correlation levels with external measures are not clear. For example, if a new generative AI-based item type purports to be a better measure of a construct because it eliminates unknown background context, we might expect a lower correlation to existing measures. An open question for discussion in the field then becomes: how do we demonstrate an innovative assessment is actually a better measure of a construct than an existing assessment?

The potential lower correlations will also present a challenge to adoption of new forms of formative assessment as long as schools and districts are held accountable through their scores on traditional assessments. Decision-makers in schools will want assessments that predict whether students are on track to be successful on end-of-year assessments even if that end of year assessment is less perfect.

Do We Need Summative Assessments?

Given the relatively large amounts of data about student performance coming from interactions with digital learning environments, some have asked whether we need summative assessments. In fact, John Behrens and I have laid out a vision for the future in the "digital ocean" where, because we have so much data from daily learning interactions, we do not need to ask people to stop and take a test (DiCerbo & Behrens, 2014). However, we are not at that place at the moment. The data collected by Khan Academy is vast; there is a large data lake full of student

interaction data. However, there is a lot of noise in the data. Students start working on a problem and walk away, then return and skip to other exercises on other skills. Student choice and agency was built into the platform on purpose to allow students to pursue individual interests. Students may be working together to solve problems with their peers (which is acceptable in a learning context). In a recent classroom visit, students were observed working in pairs and using one student's Khanmigo account to ask questions when they got stuck. Context information of this kind is not gathered on the platform and while it could be, inserting points of friction in the experience, for example, requiring students to enter names of those they are working with, decreases the likelihood of actually engaging in the learning activities. Students begin conversations with Khanmigo but then drop off, maybe because they get the help they need but perhaps because they didn't get the help they needed. For purposes of formative assessment, where the decisions being influenced are around what should be taught the next day, and there are teachers and parents in the loop to make adjustments if what is indicated by the assessment is a little off, this noise is acceptable. However, if more consequential decisions were to be made, with less chance of correcting for error, these measures are likely too unreliable in their current state to be fit for that purpose.

Concluding Thoughts

The use of generative AI to solve some of the long-standing problems in assessment sounds quite promising (or perhaps quite daunting) and there is great potential, but there is also much research to be done before these models can be used in higher stakes assessment. Even for formative assessment, a big challenge comes from the fact that large language models are, by definition, probabilistic. The responses the model gives, even from the same instruction and student input, vary each time the model produces a new response, which impacts standardization, but it could also impact the extent to which the model prompts students for more information or gives help or hints. Models can do well nearly all the time but occasionally give an odd response. In low-stakes environments, with a teacher available, wrong or illogical responses can be addressed but it would be a significant concern in higher stakes situations. More work is needed before generative AI-based tasks or scoring can be validly and reliably used for high stakes decisions.

More generally, existing digital learning experiences offer learners the possibilities of nearly unlimited practice with immediate feedback. The amount of information gathered from these opportunities is sufficient to inform instructional decisions about what students need support on, where they are succeeding, and what they should work on next. The systems are ideal for setting expectations for what is to be learned over time and providing students with feedback in ways that support learning. The introduction of generative AI offers the ability to improve on the use of information from assessments to inform instruction and also to build equitable experiences where students are not penalized for construct-irrelevant differences in background knowledge. Much of the ability to provide instructionally relevant information comes from the fact that these data are gathered over time, providing the ability to capture multiple instances of students solving problems at the skill level. At the same time, information gathered during informal practice also results in significant noise in the data, which cautions against its use in high-stakes decisions. Ultimately, data from student experiences on one platform will never capture the sum of all they know and can do, but it can help give us more information about students at a more granular level if used with care.

References

Appleton, A. (2024, June 17). Indiana schools embrace AI, but seek to 'keep humans in the loop' *Chalkbeat*. https://www.chalkbeat.org/indiana/2024/06/17/students-use-ai-pilot-programs-in-class/

Baker, E., Dickieson, J., Wulfeck, W., & O'Neil, H. F. (Eds.). (2017). *Assessment of problem solving using simulations*. Routledge. https://doi.org/10.4324/9781315096773

Behrens, J. T., Collison, T. A., & DeMark, S. (2008). The seven C's of comprehensive online assessment: Lessons learned from 36 million classroom assessments in the Cisco Networking Academy Program. In L. A. Tomei (Ed.), *Online and distance learning: Concepts, methodologies, tools, and applications* (pp. 2578–2592). IGI Global. https://doi.org/10.4018/978-1-59904-935-9

Behrens, J. T., DiCerbo, K. E., & Foltz, P. (2019). Assessment of complex performances in digital environments. *Annals of the American Academy of Political and Social Science, 683*, 217–232. https://doi.org/10.1177/0002716219846850

Castellano, K., Hoffman, E., Bauer, M., Bertling, M., Kitchen, C., Jackson, T., Oranje, A., DiCerbo, K., & Corrigan, S. (2015). *Game-Based Formative Assessment for Argumentation: Mars Generation One: Argubot Academy* [Conference presentation]. Annual meeting of the American Educational Research Association, Chicago, IL.

Choi, J., DiCerbo, K., Ventura, M., Lai, E., Wood, J., & Iverson, J. (2019). *Measuring proficiency using interactive simulation data: Empirical comparison of evidence aggregation methods* [Conference presentation]. National Council on Measurement in Education Annual Meeting, Toronto, Ontario, Canada.

DiCerbo, K. E., & Behrens, J. T. (2014). *The impact of the digital ocean on education* [White paper]. Pearson. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/DigitalOcean.pdf

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations of ongoing assessment. In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy.* Springer (pp. 1–21). https://doi.org/10.1007/978-3-319-17727-4_66-1

Foster, N., and M. Piacentini (Eds.). (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en.

Gobert, J. D., & Sao Pedro, M. A. (2016). Digital assessment environments for scientific inquiry practices. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 508–534). Wiley. https://doi.org/10.1002/9781118956588.ch21

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495–522. https://doi.org/10.1002/acp.2350090604

Grimaldi, P. (2023, November 16). Multiple studies show Khan Academy drives learning gains: Evidence for our platform's effectiveness. *Khan Academy Blog* https://blog.khanacademy.org/multiple-studies-show-khan-academy-drives-learning-gains-evidence-for-our-platforms-effectiveness/

Guskey, T. R. (2022). *Implementing mastery learning.* Corwin Press.

Kulik, C. L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265–299. https://doi.org/10.2307/1170612

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1), i-29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Morrison J. R., Ross S. M., Cheung A. C. (2019). From the market to the classroom: How ed-tech products are procured by school districts interacting with vendors. *Educational Technology Research and Development, 67*(2), 389–421. https://doi.org/10.1007/s11423-019-09649-4

OECD (2017). *PISA 2015 results (Volume V): Collaborative problem solving*, PISA. OECD Publishing. http://dx.doi.org/10.1787/9789264285521-en

Oreopoulos, P., Gibbs, C., Jensen, M., & Price, J. (2024). *Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence* (No. w32388). National Bureau of Economic Research. https://doi.org/10.3386/w32388

Reich, J. (2020). *Failure to disrupt—Why technology alone can't transform education*. Harvard University Press. https://doi.org/10.2307/j.ctv322v4cp

Senko, C. (2019). When do mastery and performance goals facilitate academic achievement?. *Contemporary Educational Psychology, 59*, article 101795. https://doi.org/10.1016/j.cedpsych.2019.101795

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge. https://doi.org/10.4324/9780203122761

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. https://doi.org/10.4324/9781410605931

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 487662. https://doi.org/10.3389/fpsyg.2019.03087

Yamkovenko, B. (2023, September 25). Why Khan Academy will be using "skills to proficient" to measure learning outcomes (and you should too!). *Khan Academy Blog*. https://blog.khanacademy.org/why-khan-academy-will-be-using-skills-to-proficient-to-measure-learning-outcomes/

Zapata-Rivera, D. (Ed.), (2019). *Score reporting research and applications*. Routledge. https://doi.org/10.4324/9781351136501

Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 585–602). Routledge. https://doi.org/10.4324/9780203102961

Game-Based Assessment: Practical Lessons from the Field

Jack Buckley and Erica Snow

Abstract

In this chapter we discuss a particular application of digital games for learning: game-based assessment (GBA). This approach to assessment allows for the measurement of a broader range of skills (e.g., "durable" skills such as creative problem solving and collaboration), as well as better measurement of some aspects of the "thinking" of respondents, including in traditional domains like science and mathematics or adult learning in the workplace. While promising, GBA is not without practical challenges. For example, game-based assessments can often be more costly and difficult to develop than traditional standardized tests based on a series of discrete questions or small "testlets" or tasks. Despite this challenge, GBA is not infeasible or impractical; in fact, we have been developing GBAs for education and workplace applications for over seven years, including in the high-stakes workforce selection context. Here we draw from our hard-earned experience in this domain and share some lessons we have learned that may be helpful for the next wave of GBA developers.

Authors Note

We would like to thank our current and past colleagues at Roblox, Imbellus, and Mckinsey & Co. who contributed to the work presented in this chapter.

Introduction

Games and learning have long been intertwined. While perhaps the earliest evidence of the use of games as a teaching tool dates at least to Classical Greece, if not to the creation of African board games some 5,000 years ago (Hellerstedt & Mozelius, 2019), the advent of digital computing marked the beginning of a new era of computer games and simulations in the service of learning.

The earliest digital learning games, such as "The Sumerian Game," developed for the IBM 7090 in 1964 (Wing, 1967) allowed learners to interact with and learn the principles of complex systems in a novel and engaging way, albeit handicapped by the technological limitations. In the subsequent decades, every advance in computing technology (e.g., home microcomputers, CD-ROM drives, the Internet, high-speed broadband, machine learning, educational data mining) have been harnessed almost immediately for learning. Simultaneously, the applications of these technologies spread across many domains and populations, from preschool mathematics to computer programming in the workplace.

Although this history is fascinating and holds many lessons for the educational content developer of today, in this chapter we concern ourselves with a narrower subset of the application of digital games for learning: *game-based assessment* (GBA). This approach to assessment allows for the measurement of a broader range of skills (e.g., "durable" skills such as creative problem solving and collaboration), as well as better measurement of some aspects of the "thinking" of respondents, including in traditional domains like science and mathematics or adult learning in the workplace.

While promising, GBA is not without practical challenges. For example, game-based assessments can often be more costly and difficult to develop than traditional standardized tests based on a series of discrete questions or small "testlets" or tasks. Despite this challenge, GBA is not infeasible or impractical; in fact, we have been developing GBAs for education and workplace applications for over seven years, including in the high-stakes workforce selection context. In the pages that follow, we will draw from our hard-earned experience in this domain and hopefully share some lessons we have learned that may be helpful for the next wave of GBA developers.

The remainder of this chapter is organized as follows: after a brief discussion of some preliminaries and definitions, we turn to a description of our GBA design process. We then illustrate that process with several real examples from our work at both Imbellus, a GBA startup, and Roblox, a gaming platform technology company. We share examples (and lessons) from both the K–12 education and workforce learning contexts. We conclude with some thoughts on the future of GBA.

Preliminaries

Why Game-Based Assessment?

In our experience there are two primary reasons to consider the development of a GBA instead of taking a more traditional (and often less costly) approach. The first is that, compared to traditional assessment, GBA can allow for *measuring different constructs*. Increasingly, in both P-20 education and in workforce learning and selection, there is significant interest in measuring "durable skills" (or "soft skills" or "21st Century Skills") such as critical thinking, communication, computational thinking, collaboration, systems thinking, and creative problem solving (Trilling & Fadel, 2009). The use of games or simulations (more on the distinction below) is a promising way of measuring these constructs (Stecher & Hamilton 2014; Seelow 2019).

Aside from durable skills, curricular frameworks in P-20 education around the world are increasingly multi-dimensional and include cross-cutting skills as well as traditional academic content. For example, the Next-Generation Science Standards (NGSS Lead States, 2013) in the United States include scientific practices and cross-cutting concepts as well as traditional scientific domain knowledge. These new dimensions can be difficult to assess via traditional means (Smith et al., 2022). As global education systems increasingly expand their curricular standards to include these kinds of constructs, there will be increasing demand for formative and summative assessments to keep pace.

The other reason to consider GBA is that the use of games allows the test developer to *measure constructs differently*. Even if one's task is to assess learners' knowledge of familiar and relatively uncomplex content such as traditional mathematics, vocabulary, or factual knowledge, the use of GBA can improve engagement and immersion (Hamari et al., 2016). This increased test-taker engagement can be particularly important in applications like pre-hire workforce assessment, where candidates are not a "captive audience" and can simply choose to exit the application process.

However, regardless of the domain, it is important to remember that GBA is not a panacea for differences in opportunity-to-learn. If learners do not have equal access to instruction in the basic building blocks of a given domain, layering a game into the assessment experience will not ameliorate this (Porter 2007). It is also worth noting that games played for enjoyment do not have to meet the test-maker's criteria of validity and reliability. GBA, while more engaging and immersive than a "bubble sheet" test, it is constrained in many ways (Oranje et al., 2019).

Game-Based vs. "Gamified"

In recent years the idea of "gamification" or the layering of game-like elements (e.g., leaderboards, badges, or personalized avatars) to non-game educational and assessment content and tasks (Deterding et al., 2011) has become pervasive. This practice may, indeed, increase learner engagement, but we draw a distinction between this gamification and the development of true games for learning and assessment. Citing Salen and Zimmerman's (2004) definition of a "game" as, "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome," Plass, Homer, and Kinzer (2015) provide an example that illustrates the distinction between games and gamification:

> *Consider as an example the gamification of math homework, which may involve giving learners points and stars for the completion of existing activities that they consider boring. Game-based learning of the same math topic, on the other hand, even though it may also include points and stars, would involve redesigning the homework activities, using artificial conflict and rules of play, to make them more interesting and engaging.* (Plass, Homer, & Kinzer, 2015, p. 259).

We appy the same distinction for the specific case of GBA, although it is not always easy to observe in practical application.

Games vs. Simulations

Finally, it may be useful to attempt to draw a similar distinction between games and various types of "simulations." While we are not aware of any broadly-accepted definition, the typology of Narayanasamy et al. (2006) is a useful one. They distinguish between "games," "simulation games," and, "training simulations." While the three have many aspects in common, there are two important distinctions among the categories. The first is in the area of goal-orientation. Simply put, games

and simulation games are centered around goal-oriented activity, while training simulators are not. Further, games have an end state, while simulation games and training simulators continue without a determined end point (i.e., one does not "win" at Microsoft Flight Simulator).

The second area of difference among the categories is the presence or absence of a gameplay "gestalt," or pattern of interaction (perception, cognition, and motor performance) that allows for successful play (Lindley 2002). Games and simulation games both have patterns that allow for the creation of gameplay gestalts; training simulations have standard operating procedures that are well-defined and generally do not change.

Our GBA work generally seems to fall in the space between games and simulation games. The GBA tasks we have developed are goal-oriented (test-takers must complete various tasks that are transparent and quantifiable, although there are other item scores generated by their interaction with the game, as we discuss below) and allow for the formation of gameplay gestalt via patterns of perception and cognition.

Designing GBAs

The Use of Evidence-Centered Design

To develop our GBAs we use a modified version of Evidence-Centered Design (ECD; Mislevy, Almond, & Lukas, 2003), a well documented and validated approach to task design that has been used across a variety of domains and media (Frezzo, Behrens, & Mislevy, 2010; Liu & Haertel, 2011; Sweet & Rupp, 2012).

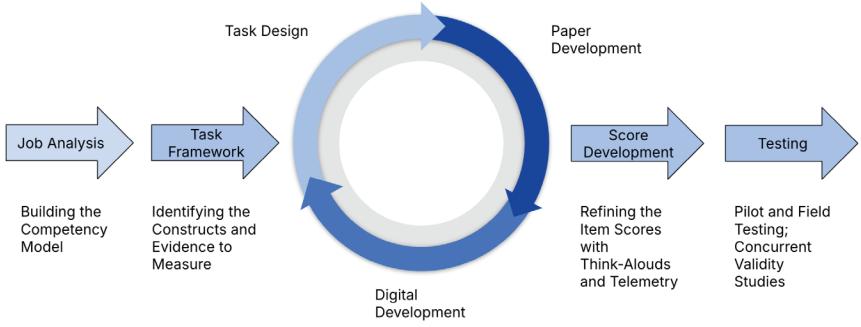
Our GBA development starts by identifying the constructs or KSAs (Knowledge, Skills, and Abilities) of interest. We identify these constructs or KSAs through cognitive task analysis or job analysis, which identifies the underlying skills, thinking, and abilities required to successfully perform a task and/or demonstrate a standard of knowledge. For hiring selection assessment these skills are often identified as key indicators of success at the company within the specific role.

Once we have conducted the job analysis and identified the target constructs/KSAs we begin to develop a task framework which will be used as a starting point for developing our GBAs. These frameworks help facilitate the collaboration between game designers, learning scientists, content experts, data scientists, and

psychometricians by identifying 1) the primary KSAs that we as scientists and designers want to build the task around, 2) the specific pieces of evidence that need to be collected to capture the KSAs, and 3) the constraints and structures potential game-based tasks must include.

After our scientists and content experts develop a task framework, we bring in our game designers and UX/UI experts for iteration on creating possible GBA tasks that meet the requirements outlined in the task framework. Our scientists walk the design team through the task framework with a specific focus on the evidence we need to collect within a possible task. Then the design team begins to iterate on possible narratives/scenarios that could be used to build out the task. As we begin to map out the various task designs we start a prototyping process that begins with paper prototypes and then shifts to digital prototypes as the work progresses. We conduct think-alouds (sometimes called cognitive labs) to gauge both usability issues with the possible tasks as well as "pressure test" the assumptions we are making about the types of thinking the task evokes and requires for successful completion.

Developing the GBA Tasks: A Modified ECD Approach



Stealth Assessment and Scoring

To score users' performance within our game-based tasks we take a stealth assessment approach to scoring (Shute, 2011). Stealth assessment provides an unobstructed view into the cognitive process of the user while they engage in the GBA. The user does not know what they are being scored on and, in most cases, it is not immediately obvious what is being measured. This allows for a more authentic view of their skills and abilities. We build our stealth assessments using the designed telemetry data generated by interaction with the task. That is, every item score is computed using test-takers' telemetry within the task. Telemetry captures the test-takers' every choice, behavior, timestamp, and click within the GBA. Every item score is pre-developed through the modified ECD process, not based on a "black box" modeling approach.

Development of item scores is a meticulous process that requires our interdisciplinary team to outline out how each potential behavior (or patterns of behaviors) maps to a specific construct and how that behavior can be transformed into an item score. Once an initial set of items is identified, we build preliminary pseudo-code for each of these items. This pseudo-code specifies algorithmically how different behaviors will be scored using the telemetry data generated by the actions players engage in the GBA. Item scores are tested throughout the prototyping process and at a full pilot stage. Data is collected and the team monitors overall item performance and construct coverage.

Evidence Centered Design (ECD) and stealth assessment provide frameworks for finding evidence of knowledge, skills or abilities in game-based assessments. This approach also can assist in combating cheating as it is not immediately clear within the game what the "right answer" is and often, there are many correct answers or ways that an item can be scored to give the test-taker full credit. This assessment approach within games allows an unobstructed look at a series of evidence identifying not only what a user knows, but the process they engaged in to get there.

Design Challenges

One of the biggest challenges in developing GBA is its interdisciplinary nature. While all cognitive assessment is (or should be) interdisciplinary to some degree (Pellegrino, Baxter, & Glaser 1999), successful development of GBA requires an exceptionally broad range of domain and disciplinary participation, including Learning Science, User Interface/User Experience (UI/UX) Design, Game Design, 3D Art, Software Engineering, Psychometrics, and Data Science (Table 1).

Table 1.
A Typical GBA Development Team

Role	Quantity
Overall Lead	1
Project Manager	1
Learning or Cognitive Scientist	1–2
Industrial-Organizational Psychologist (workforce) or Content Expert (education)	1
Game Designer	1
3D Artist	1
Data Scientist	2
UI/UX Designer	1
Game Development Lead	1
Game Developer	1–2
Backend Engineer (if integrations required)	1
Psychometrician	1

No one discipline owns the entire process; instead there is a series of hand-offs throughout the development cycle that require high levels of attention to detail and constant communication. While our learning scientists kick the process off through construct identification and development of the design pattern, the first major handoff is to a game design team. This design team may or may not initially have experience in game-based assessments and what works in the world of game design for entertainment does not always work for assessment. As the designers build out a narrative, the data scientists and psychometricians need to have constant eyes on the design to make sure the evidence needed to develop item scores is included.

Often the design team will want to have flawless user experience in the UI/UX phases, however, that may result in poor measurement. For instance, when designing a guidebook for a task, from the UI/UX perspective it is a better user experience to have fewer clicks or choices to be able to access information, resulting in less friction for the player. However, for measurement we want to include added clicks and actions to be sure exactly what a user is looking at and how they decided to access that information. This can result in added layering or nesting of information.

These differences in philosophies often put disciplines at odds. Thus, iteration is present throughout the entire process from early design all the way to operational testing. This type of interdisciplinary work requires flexibility with everyone keeping an eye on the common goal, building a reliable and valid assessment. This goal can sometimes come in conflict with other goals such as user engagement, enjoyment, and experience.

Digital GBA at operational scale also requires an entire software engineering team, consisting of game developers and, possibly, backend engineers if the game-based task must be integrated into other reporting or analytics systems. Once again, until this team gains experience with peculiarities of GBA (compared to entertainment game development), there will likely be friction between them and the assessment science professionals.

Why Don't We Just Use Existing Games?

If designing game-based assessments is such an interdisciplinary challenge, why not simply adapt existing commercial (or academic) games for measurement in the classroom or workforce? Certainly performance on some existing games is correlated with the sorts of cognitive and durable skills we seek to measure. For example, Simons et al. (2021) show that business school students with higher scores on the award-winning commercial strategy game Civilization, "had better skills related to problem-solving and organizing and planning than the students who had low scores."

While we believe there could be some efficiencies in using existing games as assessment, we have four major concerns with this approach, especially in the high-stakes context:

1. Fairness: existing games are generally designed to be entertaining, not to ensure that all test-takers have an equal opportunity to demonstrate KSAs/ competencies;

2. Content alignment: existing games are unlikely to be designed to allow evidence statements based on curriculum designers, employers' (or others') required competencies.

3. Construct-irrelevant variance: commercial games often have interlocking game systems and design elements that are uncorrelated with the constructs of interest and may be extremely distracting;

4. Time: amount of time available for selection at the top of a hiring funnel (or even in a college entrance examination) is limited compared to the time spent playing many existing games, so it can be difficult to generate item scores efficiently.

For these reasons, we generally advise teams building GBA to design their own experiences using a principled process like ECD.

Fairness and GBA

One of the guiding principles for all assessment is fairness. As the sixth Principle of this Handbook states, "Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences." (Baker et al., 2025). The premise of testing is that

tasks provide evidence of skill mastery for all examinees. If any factors unrelated to skill affect performance, assessment validity is diminished. Indeed, according to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 49), "fairness is a fundamental validity issue." In addition to the typical fairness areas of concern to all test makers, GBA introduces additional complexities. Chief among these is the need to ensure that background knowledge of and experience with games and gaming does not provide an unfair advantage to the test-taker.

One way to ensure that gaming experience does not create inequity is to measure test-takers' experience with games and conduct the same sorts of group difference and differential item functioning (DIF) analysis that one would usually conduct on sociodemographic categories like gender or primary language of instruction (or in the workplace). For example, in our work, we frequently capture the self-reported video game experience of our test-takers and construct a reference group of infrequent gamers (e.g., less than 10 hours played in the last 12 months) and a focal group of more frequent gamers. We then estimate quantities like item-level DIF, percent correct by group, and scale scores by group (including interactions with other sociodemographic factors) to ensure that we observe no substantively significant differences. If we detect DIF or see large group differences, we redesign item scores or even aspects of the GBA task as necessary to ameliorate.

It is worth noting that game experience or familiarity does not always theoretically predict better assessment performance on GBA. One reason for this, which we have seen in practical application, can be explained by the aforementioned idea of gameplay gestalt (Lindley 2002). Simply put, very experienced gamers may develop ingrained perspectives about gameplay and possible game-states due to repeated play of other games. This can cause these test-takers to make incorrect assumptions about the GBA tasks by relying on this experience to categorize them, possibly leading to the use of suboptimal heuristics instead of appropriate cognition. If this effect is detected in testing, the GBA task may require substantial redesign.

Finally, another way of ensuring fairness of GBA for non-gamers is the familiar strategy of creating and disseminating test guides and practice materials—including actual playable practice GBA tasks to help familiarize non-gamers with the user interface and "feel" of game-based assessment and, as we discuss below, reduce test anxiety.

Validity and GBA

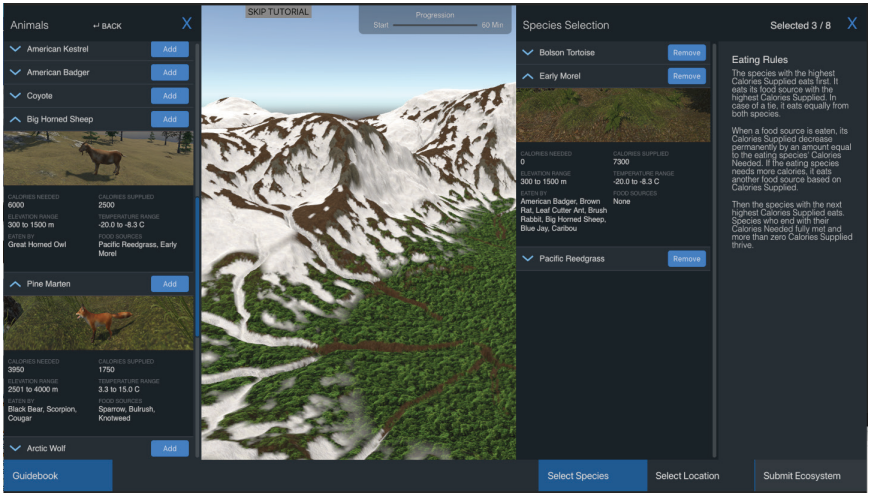
Beyond the important dimension of fairness, developers of GBA must build a broader validity argument supporting particular uses of their assessments in the classroom or workplace. As in the case of traditional assessment, this argument must cover the breadth of validity research, including but not limited to face validity, content and construct validity, concurrent and predictive validity, and consequential validity (Ferrara et al., 2016). Since GBA may be novel to both test-takers and classroom or workplace decision makers using the results, some types of validity may be challenging but important to demonstrate. We highlight some specifics in the examples below.

Examples of GBA: Imbellus

Before coming to Roblox, our team worked at a small GBA startup, Imbellus. Using the processes and techniques outlined above, we developed a hiring assessment to select new business analysts for the global consultancy, McKinsey and Company. For this assessment we had two primary tasks that were operational and part of the selection process: Ecosystem Placement (EP) and Pathogen Spread (PS). Both tasks were designed to measure cognitive skills that had been shown to be important for success at McKinsey.

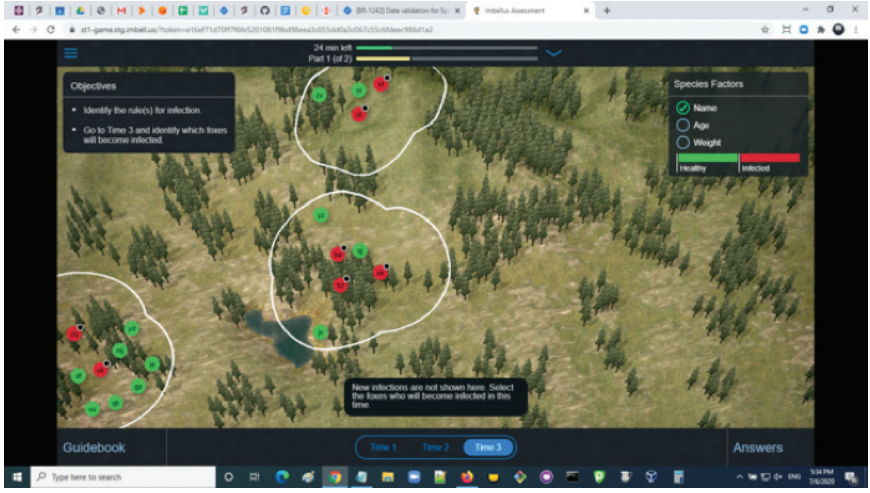
The Ecosystem Placement task measures test-takers systems thinking and situational awareness. In this task, test-takers are presented with a 3D landscape and given the goal to create a sustainable ecosystem within that environment. Test-takers are given a list of possible species that they can use to build out their ecosystem. Each species has caloric needs, environmental requirements, and predator-prey relationships that they must consider as they engage in the task.

Figure 1.
A screenshot of the Imbellus Ecosystem Placement Task.



The Pathogen Spread task measures test-takers' situational awareness and reasoning ability. In this task, test-takers are presented with a scenario where a pathogen is spreading through an animal population. Test-takers are given the goal to predict the pattern of the pathogen based on evidence given to them within the scenario such as animals' infection statuses across time, space, and other variables such as age, weight, and temperature.

Figure 2.
A screenshot of the Imbellus Pathogen Spread Task.



The Validity Argument for Ecosystem Placement and Pathogen Spread

While specific details of the validity research supporting the use of these tasks in hiring at McKinsey must remain confidential, we can provide an overview of the framework of the overall validity argument. Briefly stated, the argument demonstrates that:

- The assessment content is based on skills required by the job;
- The GBA tasks demand that players demonstrate these skills;
- This use of skills is observable and scored appropriately;
- The assessment structure reflects target content coverage;
- Recruiters are able to interpret and use assessment scores to make appropriate decisions;
- Applicants perceive the tasks as measuring relevant skills at the appropriate level of difficulty, and
- Scores on the assessment are associated with concurrent and predictive measures of candidate quality.

On the last point, during development, pilot/field testing, and operations, we were able to demonstrate concurrent validity through domain expert/novice contrast, correlations with existing instruments measuring at least part of the same domain (systems thinking, situational awareness, deductive reasoning), and predictive validity through comparing GBA performance to hiring outcomes and early job performance.

Adapting for Education

In 2019, we began to expand into the educational space by developing an adaptive, game-based assessment focused on life science content and science standards. PEEP—Project Education Ecosystem Placement was a staged adaptive GBA task aimed at measuring and providing feedback on problem solving processes for K–12 learners. Within PEEP, test-takers were asked to construct sustainable ecosystems based on the constraints of the game-based environment. PEEP was funded by the Walton Family Foundation, and was adapted from the original ecosystem placement test developed for McKinsey.

Unlike the industry version, PEEP was adapted to be more aligned and reflective of accurate life sciences content taught in schools, particularly a subsection of the Next Generation Science Standards (NGSS Lead States 2013). It was also designed to be developmentally appropriate for secondary school-aged children and also integrated elements of accessibility that would be necessary for it to be used in a school setting. PEEP was initially designed to be used as a high-stakes, summative assessment that adapted to the student's skills as they engaged with the task. PEEP was modular, where students would be asked to build out multiple ecosystems across varying environments. Each module would vary in its levels of difficulty and complexity. Complexity and difficulty would be scaffolded based on the students' performance in the previous module.

Piloting PEEP

We piloted the PEEP task in late 2019 with students from 8th to 10th grade at various school districts across the United States. Two studies were conducted to better understand students' and teachers' perceptions of the task, underlining scoring distributions. Information gathered from these studies was used to iterate and further improve the PEEP assessment task.

First, we conducted think-aloud studies where students would play through the task and, as they did, they would be prompted to describe what they were doing, why they were doing it, and their experiences with the game interface. Results revealed that students found the task enjoyable, engaging, and relevant to what they were learning in school. Interestingly, the younger students expressed more interest and engagement in the task compared to the older students, however both groups had overall positive sentiment. Teachers found the task engaging and a fun supplement to add to their curriculum. However, teachers did express concerns about the GBA's alignment to Next Generation Science Standards. They also had reservations about the scoring, interpretations, and reporting functions of the task

After think-aloud testing, we also conducted a small scale pilot where students went through the PEEP task at their own pace. This was done in classrooms and without researchers or teachers asking the students to explain what they are doing or why. This simulates a test taking environment for the student to give us more accurate data. Similar to the think aloud findings, results from this study revealed that over 80% of students who engaged in the task expressed positive sentiment towards it and felt it was relevant to their school work. Initial results showed that the underlying scoring for PEEP was working and showing variance in score distribution across students.

While these results were promising, PEEP was never implemented in schools beyond this initial work. In 2020, Imbellus was acquired by Roblox and the team transitioned toward working on the Roblox platform to develop hiring assessments as well as contribute to the educational community that is growing at Roblox.

Examples of GBA: Roblox

The "Roblox Problem-solving Assessment" (PSA) is a GBA designed to evaluate the problem-solving competencies of applicants for a variety of technical positions at Roblox, a US-based digital gaming platform technology company where the authors work. Our hiring assessments are developed and tested specifically for Roblox and the needs of our workforce. The assessment development and testing process are guided by rigorous scientific frameworks and best practices from the fields of Learning Science, Psychometrics, and Data Science. The use of an automated, standardized assessment provides an equitable opportunity for all candidates, regardless of background, to demonstrate job-relevant skills.

Roblox chose to develop GBA for hiring selection for both of the reasons cited above: measuring different constructs and measuring familiar constructs differently. First, the Roblox PSA is designed to ensure that each candidate is given the opportunity to demonstrate critical skills and abilities that are important to their prospective role at Roblox. These include hard-to-measure competencies like systems thinking and creative problem-solving, which are amenable to GBA. Second, even for some target constructs that have non-GBA, off-the-shelf assessments available (e.g., aspects of personality and computer coding ability), Roblox wanted an engaging assessment that showcases its own technology as part of the hiring process—hence GBA.

Construct Identification

The first step in developing the Roblox PSA occurred in 2021, when we identified the constructs necessary for success in the roles of interest. To accomplish this our psychologists conducted a broadly-scoped job analysis, including over 100 interviews with Engineer and Product hiring managers and leaders and collected data and artifacts on their job duties. During these interviews, respondents identified KSAs that are targeted during the selection process, important for success at Roblox, and that distinguish experts from novices across various roles. The major themes across the interview responses were summarized for both junior and senior roles across the Engineering and Product functions.

The identified KSAs were then ranked as most viable for a game-based medium using a literature review and whether or not the KSA is already being measured as a part of the hiring process. There were four categories of KSAs or competencies identified: cognitive, intrapersonal, interpersonal, and practical. Based on a literature review, market research, and the signals already being collected during the interview process, we decided to develop two game-based tasks that focus on key cognitive skills and abilities of applicants, which we built using the Roblox game engine and platform over two years using the ECD approach described above.

When evaluating candidates for roles at Roblox, we are interested not only in strong technical ability, but also in the application of those skills and abilities during complex cognitive processes. Our job analysis demonstrated that complex skills such as creative problem solving and systems thinking are necessary for success in the target roles, and high levels of ability in these areas indicates potential to make a long-term positive impact at the company. There are currently two tasks

that are in-use operationally: "Robots" and "Factories." The Robots task is designed to measure creative problem solving, specifically ideation and divergent thinking. The Factories task is designed to measure systems thinking skills. Both creative problem solving and systems thinking were identified as critical skills for success based on an extensive job analysis done between 2020–2022.

Figure 3.
Factories Task within Roblox.

The Roblox GBA uses the same development and scoring techniques mentioned above. Within the Roblox GBAs performance is measured based on patterns of behaviors that applicants exhibit within the task while they engage in the problem solving process. Generally, there are no "right" or "wrong" answers like one would see on a traditional test. Instead, we look to quantify how they get to a solution and the steps they took to get there. These item scores are not based on machine learning or black box techniques, using ECD (Mislevy et al., 2013), we outline the items during task development so we know what actions a player can take and how we will develop a scoring code around various patterns.

Validating the Roblox Problem-Solving Assessment

Similar to our work with McKinsey, we have continued to develop a program of research leading to a multi-faceted validity argument supporting the use of our GBA tasks for hiring at Roblox. The framework of this research has been largely the same, ranging from measuring the face validity of the tasks through applicant survey to comparing scores concurrently with external non-GBA measures of the same or similar constructs (creative problem solving, systems thinking), to prediction of candidate quality and performance (correlation with expert-scored resumes, prediction of performance at later stages of the hiring process, prediction of performance on-the-job).

Reducing Anxiety: Roblox Practice Test

There is a large literature in assessment extolling the virtues of practice tests as a key part of assessment (e.g., Adesope et al., 2017 for a review in the education context). Allowing test-takers to engage with test content and format can reduce anxiety and improve measurement validity. At Roblox, a key component of the use of GBA has been to also provide an opportunity for applicants to familiarize themselves with the Roblox PSA environment, especially the UI/UX aspects of a GBA, which might be unfamiliar to some candidates.

In 2023, we launched "Kaiju Cats," our practice GBA task that encourages candidates to familiarize themselves with game-play elements used in the hiring assessments in a pressure-free environment. The goal of this tool was to provide candidates with an easy (and stress-free) way to get familiar with the test format and reduce test anxiety for those who may not feel comfortable with game-based elements. Initial pilot results revealed that Kaiju Cats lowered test anxiety among applicants (through pre/post measurements), particularly for those applicants who did not have prior Roblox experience. The practice test is live on the Roblox platform and open to the public and we advertise it heavily in recruiting events as well as all applicant communications. As of late 2024, over 300,000 users have engaged with the task on the Roblox Platform.[1]

1 Anyone with a free Roblox account can try Kaiju Cats at https://www.roblox.com/games/13977123257/Kaiju-Cats.

Figure 4.
Screenshot of Kaiju Cats available publicly on Roblox.

Roblox Community Fund - Education

In 2021, Roblox created a Community Fund to provide grants to pairs of developers and educational organizations to develop new, educationally focused experiences on the Roblox platform. Many of the grant recipients were educational partners who already work with thousands of educators and millions of students across formal and informal educational settings.

Our team at Roblox supported this work by developing artifacts, tools, and acting as consultants on many of these projects. Many of the developers working in this space have limited experience with building educational games and simulations and even less experience with GBA. Our team was asked to step in and help fill the gap by building out ECD documents, leading workshops, and meeting on a regular basis to talk through measurement strategies and data collection techniques. At the end of 2024, we have contributed to 5 separate experiences that are currently live on Roblox that are accessible by students, parents, and teachers.

One of these experiences is Mission: Mars, a free educational experience available on Roblox and developed in collaboration with the Boston Museum of Science and

Filament Games.[2] In Mission: Mars, students are astronauts on Mars and have to engage in a variety of problem solving tasks while they explore the planet. Our role in supporting this work included meeting with members of both the design team and Museum of Science content team to talk about stealth assessments, proactive evidence design, and potential scoring strategies within the task.

Beyond Selection: Workforce Learning & Development at Roblox

Recently, we have begun to develop an in-house game-based conversational simulation tool as a general engine for workplace learning and development (L&D), again using the Roblox platform as the foundation. Our first use of this L&D simulation is as a way to provide our managers with training and practice delivering feedback to employees as part of a simulated employee performance review conversation—a key area of improvement identified by our internal employee listening program. The new tool provides an interactive environment to help managers practice this skill (particularly giving difficult feedback) and transfer what they learn into their actual performance conversations with employees. Similar to the assessment development and testing process, the L&D development has been guided by rigorous scientific frameworks and best practices from the fields of Learning Science, Psychometrics, and Data Science.

This game-based L&D tool specifically focuses on four areas of development for Roblox managers and leaders: how to structure a performance conversation; how to build conversations around feedback that is the most specific and relevant to the current "situation, behavior, and impact" (Bommelje, 2012); how to work with their employees to construct goals, and how to maintain supportiveness and openness throughout even difficult conversations. The primary mechanism is a series of simulated conversations with both immediate feedback to the learner (typically a Roblox people leader) after dialogue choices and end-of-conversation feedback telling the learner what they are doing right and how they could improve.

The tool is built on the Roblox platform and is designed to be easily accessible to current employees. Upon entering the task, the employee is presented with a conversational scenario, usually around giving feedback to their direct report. Employees are walked through a tutorial which outlines how to interact with the various UI elements they see during the task. The employee must complete the

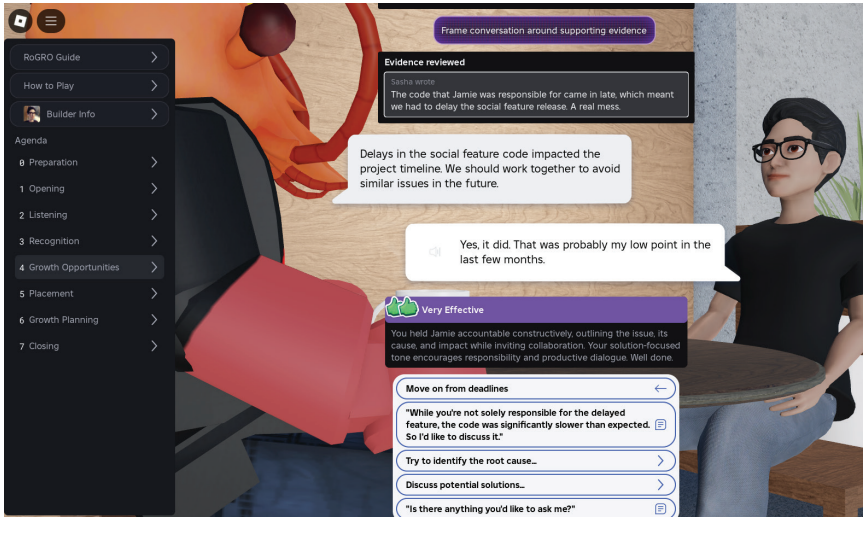
2 Similarly, anyone can try Mission: Mars at https://www.roblox.com/games/10840095864/Mission-Mars.

tutorial and then begin to prepare for the conversation with their colleague or direct report. They will use examples, peer feedback, and other evidence to support the conversation.

Once the employee enters the tool, they see their Roblox Avatar seated at a table across from a simulated direct report. Employees then begin a conversation with their direct report by selecting prompted dialogue options. Each option elicits a response from their direct report as well as real-time feedback from the tool on the effectiveness of their choice. Feedback provides the employee with areas to improve on as well as reinforces the positive behaviors they demonstrated.

As the employee progresses through the conversation, they are reminded to use evidence to provide performance feedback to their direct reports on their accomplishments and growth areas. The choices that the employee makes while engaged in the tool are recorded and scored based on their alignment to specific learning goals. After the employee exits the tool they are provided a summative report of their time in the experience and specific areas of improvements they can focus on that are tied to their performance on the learning goals.

Figure 5.
Screenshot of Roblox L&D Game.



We are currently integrating this game-based L&D simulation into our existing manager development courses, which cover a range of topics including providing manager feedback. The L&D simulation is designed to be used alongside the manager feedback course to reinforce what is being taught in the lectures and workshops through hands-on game-based practice. Managers will attend this course and then have access to the simulation where they can work through various scenarios and try out the techniques they had just learned. Practice-based learning has been shown to increase the probability of mastery compared to workshops alone (Ogrinc, 2003). This provides our managers with real-time classroom support as well as a way to practice at their own pace with the ability to return as often as they desire.

Our new prototype went live in early 2025 and we are currently conducting a series of validity studies to make sure that the game-based simulation is engaging, relevant for our managers, and ultimately improves the quality and frequency of the feedback employees receive. A full study design will be implemented in 2025 that will collect employee and manager perceptions of the tool and of the quality and frequency of the conversations they are having, performance data within the tool, and overall product usage (i.e., how do managers use the tool).

If this work shows promise we will be expanding this simulation beyond feedback conversations into other areas such as structuring effective 1:1 conversations, preparing and delivering presentations, and "managing upwards." We are also exploring the use of an integrated LLM to allow for more fluidity in the conversation as well as adaptations over time.

Concluding Thoughts on the Future of GBA

Almost a decade of designing and developing GBA in both the education and workforce environments has taught us that the approach can be very challenging, but also very rewarding. Looking ahead, we believe that the use of GBA will continue to expand and become a familiar component of many testing programs as long as the field can continue to drive development costs down and improve the underlying technology.

Controlling Costs

Compared to more traditional assessment, GBA is still very expensive on a cost-per-item (or unit of information) basis. There are several reasons for this cost differential. First, as we outline above, developing GBA requires an interdisciplinary team with a broad range of skills (game design, software engineering, cognitive scientists, assessment experts, psychometricians, etc.). Some of these disciplines, like engineering, are highly in-demand in the labor market. Second, testing and development cycles are long and early stages require frequent iteration (and, often, expensive pilot and field test data collection). Third, fully-immersive game experiences are difficult to make accessible for test-takers with disabilities, requiring either new technology or the development of equivalent means of assessment. Finally, there may be hardware and bandwidth requirements for some GBA that require investment in infrastructure.

The good news is that there are a variety of innovations and strategies that can be combined to reduce GBA development costs and ensure the method is more feasible for broad adoption. First, the explosion of generative artificial intelligence in recent years, while not useful for everything its proponents claim, does appear to be very useful at producing medium-quality code, reducing engineering costs and accelerating development. As this capability continues to improve, the costs of game engineering and UI component development will continue to decrease.

Without question, the explosion of generative AI promises to increase the efficiency of GBA development and may fundamentally change the work of many of the disciplines required. However, we have yet to see the ability of current-generation tools to completely eliminate any job function entirely. One very interesting area to watch is the application of generative AI to 3D and "4D" (animated) art, an essential part of game-based assessment development. Roblox recently introduced an open-source foundational generative model, "Cube 3D," which generates 3D models and environments directly from text and, in the future, image inputs. The generated objects are fully compatible with game engines today and can be extended to make objects functional for use in GBA.[3]

3 Code is available at https://github.com/Roblox/cube and you can try an interactive demo at https://huggingface.co/spaces/Roblox/cube3d-interactive.

Beyond generative AI, there are several additional ways to lower GBA costs even further. First, developers can build extensions to existing game development platforms to support educational and workforce assessment and release them free to the broader assessment community. Our experience harnessing the Roblox platform for GBA is a proof-of-concept experiment that demonstrates the feasibility of adapting existing technology for assessment. Second, as those existing platforms and game engines improve their ability to run on low-end hardware and slow networks (something on all technology companies' roadmaps given the need to expand their customer base globally), the cost to implement GBA in educational settings will decrease. Finally, we believe that GBA developers working on the frontier of this area can help others by sharing or licensing relevant artificial intelligence and machine learning methods and novel psychometrics methods and code libraries.

Increased Formalization

We believe there is enormous potential for the development of a more rigorous science of game-based assessment, building on the century-plus of academic and industry work that has created the foundation of modern psychometrics and measurement. Particularly promising is the emerging "General Game Playing" subfield of AI research that has led to development of multiple Game Description Languages including: S-GDL (Genesereth et al., 2005), RBG (Kowalski et al., 2017), and Ludii (Soemers et al., 2022), among others.

This is analogous to the idea of *design patterns* in architecture (Alexander 1966) or software development (Beck & Cunningham 1987), with similar potential for improving the efficiency of GBA development. This improved mathematical formalization of game elements ("ludemes") could improve scoring design and cut development and testing time. For example, equating "forms" of GBA tasks is currently complicated and data-intensive; improved formalization might get us closer to equating with little or no data (Mislevy et al., 1993). Further work in this area may also make it possible to generate games rapidly for prototyping and assessment use simply by describing a limited set of variables.

Concluding Thoughts

Almost a decade of designing and developing GBA in both the education and workforce environments has taught us that the approach can be very challenging, but also very rewarding. The combination of increased interest in measuring cross-cutting or complex cognitive constructs and durable skills in education and the workforce, coupled with the desire to make assessment more engaging, suggest a growing demand for game-based assessment, despite the relatively high start-up costs and need for an interdisciplinary development team. Looking ahead, we believe that the use of GBA will continue to expand and become a familiar component of many testing programs as long as the field can continue to drive development costs down and improve the underlying technology.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701.

Alexander, C. (1966). The pattern of streets. *Journal of the American Institute of Planners, 32*(5), 273-278.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Baker, E. L., Everson, H. T., Tucker, E. M., Gordon, E. W. (2025). Principles for Assessment Design and Use in the Service of Learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst.

Beck, K., & Cunningham, W. (1987). Using pattern languages for object-oriented programs. Paper presented at OOPSLA 1987 Conference. No. CR-87-43.

Bommelje, R. (2012). The listening circle: Using the SBI model to enhance peer feedback. *International Journal of Listening*, *26*(2), 67-70.

Deterding, S., D. Dixon, R. Khaled & L. Nacke. (2011), "From game design elements to gamefulness", Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments—MindTrek '11, http://dx.doi.org/10.1145/2181037.2181040.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation: Cognition in score interpretation and use. In *A. A. Rupp & J. P. Leighton (Eds.), The Handbook on cognition and assessment: Frameworks, methodologies, & applications* (pp. 41-74). Malden, MA: Wiley.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology*, *19*, 105–114.

Genesereth, M., Love, N., & Pell, B. (2005). General Game Playing: Overview of the AAAI Competition. *AI Magazine. 26*, 62–72. https://doi.org/10.1609/aimag.v26i2.1813

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54, 170–179.

Hellerstedt, A., & Mozelius, P. (2019). Game-based learning—a long history. Proceedings of the Irish Conference on Game-based learning, Cork, Ireland. https://www.researchgate.net/profile/Peter-Mozelius-2/publication/336460471_Game-based_learning_-_a_long_history/links/5da1c24745851553ff8ad248/Game-based-learning-a-long-history.pdf

Kowalski, J., Sutowicz, J., & Szykuła, M. (2017). Regular boardgames. arXiv. https://arxiv.org/abs/1706.02462

Lindley, C. A. (2002). The gameplay gestalt, narrative, and interactive storytelling. *Proceedings of Computer Games and Digital Cultures Conference,* June 6–8, Tampere, Finland.

Liu, M., & Haertel, G. (2011). Design patterns: A tool to support assessment task authoring. *Large-Scale Assessment Technical Report*, *11*.

Sweet, S. J., & Rupp, A. A. (2012). Using the ECD framework to support evidentiary reasoning in the context of a simulation study for detecting learner differences in epistemic games. *Journal of Educational Data Mining*, *4*(1), 183–223.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., Sheehan, K. M. & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

Narayanasamy, V., Wong, K., Fung, C., & Rai, S. (2006). Distinguishing games and simulation games from simulators. *Computers in Entertainment (CIE), 4*(1), 9. https://doi.org/10.1145/1129006.1129021

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. The National Academies Press. https://doi.org/10.17226/18290

Oranje, A., Mislevy, B., Bauer, M., & Jackson, G. T. (2019). Summative Game-based Assessment. In D. Ifenthaler & Y. Kim (Eds.), *Game-based assessment revisited.* Springer.

Ogrinc, G., Headrick, L. A., Mutha, S., Coleman, M. T., O'Donnell, J., & Miles, P. V. (2003). A framework for teaching medical students and residents about practice-based learning and improvement, synthesized from a literature review. *Academic Medicine*, *78*(7), 748–756.

Pellegrino, J. W., Baxter, G., & Glaser, R. (1999). Addressing the two disciplines problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 309–355). AERA

Plass, J. L., B. D. Homer, & C. K. Kinzer. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283.

Porter, A. (2007). Rethinking the achievement gap. *@PennGSE: A Review of Research*. https://www.gse.upenn.edu/system/files/u10/Fall_2007.pdf.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.

Seelow, D. (2019). The art of assessment: Using game-based assessments to disrupt, innovate, reform, and transform testing. *Journal of Applied Testing Technology, 20*(S1), 1–16.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503–524.

Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2021). Good gamers, good managers? A proof-of-concept study with *Sid Meier's Civilization. Review of Managerial Science, 15,* 957–990. https://doi.org/10.1007/s11846-020-00389-8

Smith, P. S., & C. L. Plumley, with L. Craven, L. Harper, & L. Sachs. (2022). *K–12 Science Education in the United States: A Landscape Study for Improving the Field.* Written by P. Sean Smith and Courtney L. Plumley. Carnegie Corporation of New York.

Soemers, D., Piette, É., Stephenson, M., & Browne, C. (2022). *The Ludii Game Description Language is Universal.* https://doi.org/10.48550/arXiv.2205.00451

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan.* RAND Corporation. https://www.rand.org/pubs/research_reports/RR863.html.

Trilling, B., & C. Fadel. (2009), *21st century skills: Learning for Life in Our Times.* Jossey-Bass.

Wing, R. L. (1967). *The production and evaluation of three computer-based economics games for the sixth grade: Final report* (Report No. ED014227). Westchester County Board of Cooperative Educational Services. https://eric.ed.gov/?id=ED014227

From Evaluation to Impact: Transforming Assessment into a Tool for Learning

Michelle Odemwingie and Kimberly Cockrell

The Achievement Network, Ltd.

Abstract

This chapter examines the evolving role of assessment, moving beyond rigid measurements toward a more dynamic, learning-centered approach. Traditional assessment models often prioritize evaluation over instructional impact, but research and practice show that assessments can do more—guiding teaching, informing student learning, and strengthening instructional coherence.

The Achievement Network (ANet) works alongside schools to design assessments that emphasize transparency, alignment, and student agency. By integrating high-quality instructional materials with assessments that provide timely, meaningful feedback, ANet supports educators in making informed instructional decisions that drive student growth.

Key areas discussed include:

- **Transparency**, ensuring educators and students can understand purpose and design of assessments and act on results.
- **Coherence,** aligning curriculum, instruction, and assessment for a more structured learning experience.
- **Student-centered design**, fostering engagement and self-efficacy.

This chapter also explores challenges such as testing overload and assessment quality, emphasizing the importance of curriculum alignment and actionable insights. Case studies from Madison Metropolitan School District, Carlsbad Municipal Schools, and Honey Dew Elementary highlight how these principles lead to meaningful instructional improvement.

Ultimately, effective assessment is about supporting—not just measuring—learning, and this chapter shares insights on how ANet is designing solutions alongside schools to make that shift.

Introduction

Imagine a classroom where assessments do more than measure learning—they propel it forward. Where students see assessments as a tool for their own growth rather than a high-stakes judgment. Where teachers receive real-time insights that inform instruction, rather than overwhelming data that offers little guidance.

For decades, assessments have been treated as tools of measurement rather than instruments for learning. Standardized tests and traditional summative assessments provide snapshots of student performance—lacking the insights educators need to guide instruction. Too often, assessments are imposed on classrooms rather than embedded within the learning process, reinforcing a system that prioritizes evaluation over growth.

Achievement Network (ANet) believes assessments should do more than measure learning—they should improve it. Over the past 20 years, the organization has been at the forefront of the data-driven instruction movement, helping schools envision a system where assessments are transparent, coherent, and student-centered, providing educators and students with the timely, actionable information they need to drive instruction forward. The approach to assessment design anchors on them being tools that do more than just evaluate knowledge but also enhance the teaching and learning process.

At the heart of this approach is a belief that assessments should be most accountable to the student, enabling a dialogue between the learner and what is to be learned, revealing what has been accomplished and how far there is to go to achieve mastery. This vision is more than theoretical. Through partnerships with schools across the country, ANet has developed assessments that are embedded in instructional cycles, offering teachers the feedback they need when they need

it. By aligning our assessments with high-quality curricula and instructional best practices, we ensure that they are not just accountability measures but catalysts for deeper learning.

This chapter explores the Achievement Network (ANet) approach to assessment, the critical challenges facing educators today, and the ways in which well-designed assessments can transform classrooms. Our goal is to push the dialogue forward, challenging outdated assessment models and outlining how ANet is reshaping assessment to center learning. To do this, we start with our theory of action: a commitment to making assessments transparent, coherent, and student-centered.

Theory of Action

ANet advocates for a fundamental shift in assessment, moving from static measurement to dynamic tools that actively support learning. Our approach integrates high-quality instructional materials with assessments designed to inform and enhance instruction in real time. By embedding assessments into instructional cycles, ANet provides educators with timely, detailed feedback, allowing them to make informed decisions that directly impact student growth and achievement.

ANet's Theory of Action maintains that assessments should improve learning and the teaching cycle that supports instruction. This approach is rooted in three guiding principles:

- **Transparency:** Educators and students need clear access to items, design, and results. When teachers understand the rationale behind assessment design and can easily interpret results, they are better equipped to use data effectively.
- **Coherence:** Assessments must align with instructional goals and curricula to create a seamless learning experience. Rather than being separate entities, assessments should serve as integral tools that reinforce the curriculum and provide actionable insights.
- **Student-Centered Design:** Assessments should be fair and accessible to all students, ensuring that diverse learning experiences are accounted for. ANet prioritizes the development of assessment tools that reduce bias and support all students in achieving their full potential.

ANet emphasizes that assessment should be more than a compliance exercise. It should drive deeper learning. Our work develops assessments that foster deeper learning, support instructional adaptability, and empower students, teachers, and leaders. Drawing on research in socio-cognitive and sociocultural learning models, we design assessments that measure knowledge and promote higher-order thinking and student agency. By equipping educators with the insights needed to refine instruction, we ensure that assessments are used as instruments of progress rather than barriers to learning. By collaborating closely with school leadership teams, we enhance instructional leadership and support effective use of curricula and assessments. This holistic approach ensures that design, strategy, and implementation operate in concert, driving substantial improvements in educational outcomes.

Through this approach, ANet envisions a future where assessments are fully embedded in the learning process, driving both instructional excellence and student success. Grounded in research and real-world practice, ANet's approach demonstrates how well-designed assessments can transform education, shifting assessments from tools of evaluation to instruments of learning.

ANet's Assessment Design: Transparency, Coherence & Student-Centered Design

Transparency in Assessment: Building Clarity, Trust, and Instructional Impact

> *"I have concluded that building upon a long and extraordinary history of achievement in the assessment* OF *education, the future of assessment in education will likely be found in the emerging interest in and capacity for assessment to serve, inform, and improve teaching and learning processes and outcomes. Shall we call that assessment* FOR *education in addition to the assessment OF education?"*—Edmund Gordon (2013)

From its inception, ANet has designed assessments as an integral component of the teaching and learning process. ANet's assessment system is anchored on formative interim assessments designed to provide educators with timely, actionable insights that inform instruction. The goal is to create assessments that reflect rigorous academic standards while also offering a practical framework

for teachers to diagnose student learning needs, adjust instruction, and support student growth.

Despite the growing emphasis on high-quality instructional materials, opportunities for formal peer review of interim assessments remain limited. Unlike summative assessments, which undergo extensive evaluation processes, interim assessments are rarely subject to the same level of scrutiny and external validation. This gap makes third-party reviews, such as the Louisiana Department of Education's (LDOE) Tier One rating system, a critical benchmark for ensuring quality and alignment. At the core of ANet's assessment design is a commitment to ensuring that teachers have visibility into what students know, where misconceptions arise, and how instruction can be adapted accordingly. To achieve this, ANet has continually refined its approach, ensuring alignment with rigorous college- and career-ready standards and state-level expectations, as reflected in its Tier One rating from LDOE.

Approach to Design Structure: Content

ANet assessments are intentionally structured to balance rigor, alignment, and usability, ensuring that teachers receive meaningful data without disrupting the flow of instruction. To support instructional coherence, assessments are designed for real-time classroom use, allowing teachers to adjust instruction as needed. The structure ensures that assessments serve as a seamless part of the learning process—providing actionable insights without overshadowing instruction.

Structuring Literacy Assessments: A Text-First Approach

In English Language Arts (ELA), ANet prioritizes a text-first approach that mirrors the depth and complexity of high-quality reading instruction. Instead of isolating individual skills, the assessments are structured to evaluate students' ability to:

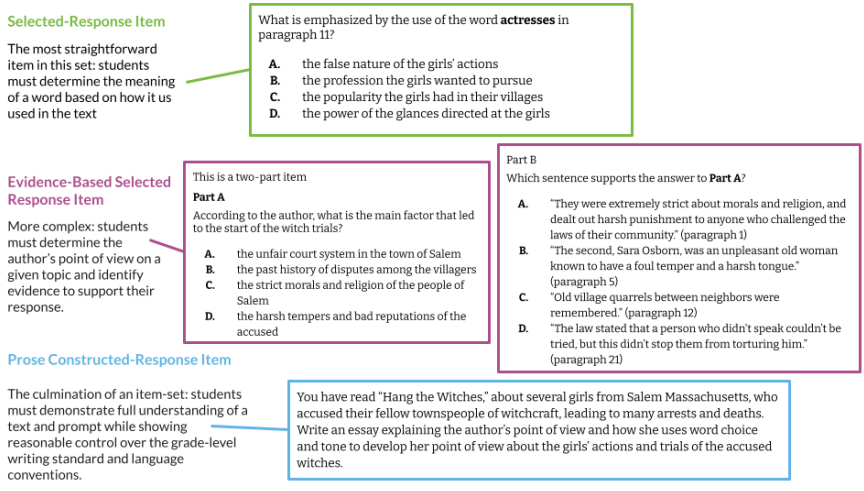
- Comprehend and analyze complex literary and informational texts
- Use text-based evidence to support reasoning
- Apply higher-order thinking to interpret and respond to questions

Key Features of ANet's Literacy Assessments:

- **Authentic Text Selection:** Over 90% of texts are previously published, high-quality selections, covering a balance of literary, nonfiction, poetry, and technical texts.
- **Standards-Based Item Development:** Each machine-scored item aligns with state level reading and/or writing standards, ensuring precision.
- **Variety of Question Types:** Questions include multiple-choice, evidence-based selected response, and constructed-response tasks that require synthesis of evidence across texts (See Figure 1).
- **Writing to Sources:** Assessments integrate tasks requiring students to analyze, compare, and synthesize ideas from multiple texts.

Figure 1.

ELA Item Sets Including Item Types and Ranges of Difficulty (7th grade)



These features earned ANet's ELA Interim Assessments a Tier One rating from LDOE, recognizing their alignment, rigor, and design. The review highlighted:

- **Text quality and complexity:** The assessments feature Lexile-appropriate texts that support deep comprehension.
- **Text-dependent questions:** Nearly all questions require direct textual evidence, ensuring students engage deeply with reading material.
- **Comprehensive writing assessments:** Students are required to craft well-defended arguments, synthesize research, and analyze literary themes, making these assessments a robust measure of college and career readiness.

This design moves literacy assessments beyond recall, emphasizing analytical thinking and engagement with complex texts.

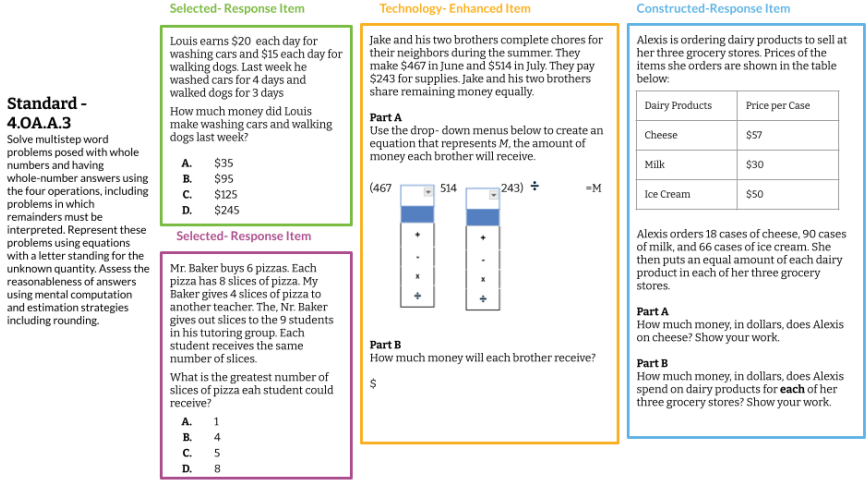
Mathematics Assessment: Balancing Rigor and Conceptual Understanding
ANet's math assessments evaluate students across three dimensions of rigor:

- **Conceptual Understanding:** Deep comprehension of mathematical principles
- **Procedural Skill & Fluency:** Accuracy and efficiency in computation
- **Application:** Applying math skills in real-world contexts

Key Features of ANet's Mathematics Assessments:

- **Aligned to state standards and Mathematical Practice Standards:** Ensuring consistency with state and national expectations.
- **Emphasis on Major Work of the Grade:** At least 65%-80% of score points target priority standards, reinforcing mastery of key mathematical concepts.
- **Innovative Item Types:** Multi-part questions, coordinate plane graphing, number line activities, and interactive technology-enhanced items assess depth of understanding (See Figure 2).
- **Misconception Analysis:** Incorrect answer choices target common misunderstandings, providing insight into student learning gaps.

Figure 2.
Math Item Sets Including Item Types (4th grade)



ANet's Math Interim Assessments earned a Tier One rating from LDOE for their strong alignment, rigor, and instructional value. The evaluation highlighted:

- **High alignment to grade-level standards:** Over 90% of test items fully reflect standard intent.
- **Balanced rigor**: Integrates conceptual understanding, procedural fluency, and real-world application.
- **Diverse item formats:** Includes multiple-choice, multiple-select, numeric response, and constructed-response tasks.

This multi-dimensional approach transforms math assessments from procedural drills into opportunities for deep mathematical reasoning.

Transparency: Designing Assessments for Instructional Alignment and Actionable Insights

Transparency in assessment is more than access to data—it is about ensuring that educators and students can interpret, understand, and act upon assessment results in ways that drive instructional improvements. When assessment data is clear, accessible, and actionable, it transforms teaching and learning, allowing educators to make evidence-based decisions and students to engage in their own learning progress. ANet's assessment design and data reporting systems are structured to provide deep visibility into student learning, ensuring that assessments both measure learning goals and can be leveraged as instruments for growth.

An Open Book Approach to Assessment Design

ANet's commitment to transparency begins at the foundational level of assessment design. Every aspect of an assessment—from the rationale behind text complexity to the reasoning behind multiple-choice distractors—is explicitly shared with educators. By offering a full window into assessment choices, ANet ensures that teachers are not only informed consumers of assessment data but also active participants in interpreting and applying results to instructional practice.

Research emphasizes that making assessment criteria explicit is essential for meaningful instructional use. Pellegrino, Chudowsky, and Glaser (2001) argue that educators and students benefit when learning goals, expected performance levels, and assessment criteria are clearly articulated. Transparent assessment materials not only support instructional decision-making but also foster a shared understanding of achievement standards among teachers, students, and the broader educational community.

To demystify assessment design and data reporting, ANet provides structured resources that equip teachers to understand, analyze, and respond to student learning needs:

- **Assessment Design Guides:** ANet provides educators with detailed rationales for text selections, Lexile levels, and question types, ensuring that assessments align with grade-level expectations and learning standards.

- **Rubric and Scoring Clarity:** Educators receive rubric interpretation guides that clarify expectations for written responses, providing explicit scoring criteria and examples to ensure consistent and meaningful assessment of student work.
- **Multiple-Choice Distractor Rationale:** Each incorrect answer choice in ANet's assessments is deliberately constructed to reflect common student misconceptions. ANet provides teachers with detailed explanations of why each distractor exists, allowing educators to diagnose student misunderstandings more effectively (Figures 3 and 4).
- **Student Work Analysis Tools:** Beyond simple correct/incorrect responses, ANet provides tools for analyzing how students approach problems, reinforcing diagnostic insights that support targeted interventions.
- **Reteaching & Instructional Support Tools:** Once educators identify areas of student need, ANet connects assessment data to actionable reteaching strategies, ensuring that every data point leads to an instructional next step.

Figure 3.
ANet Distractor Rationale ELA

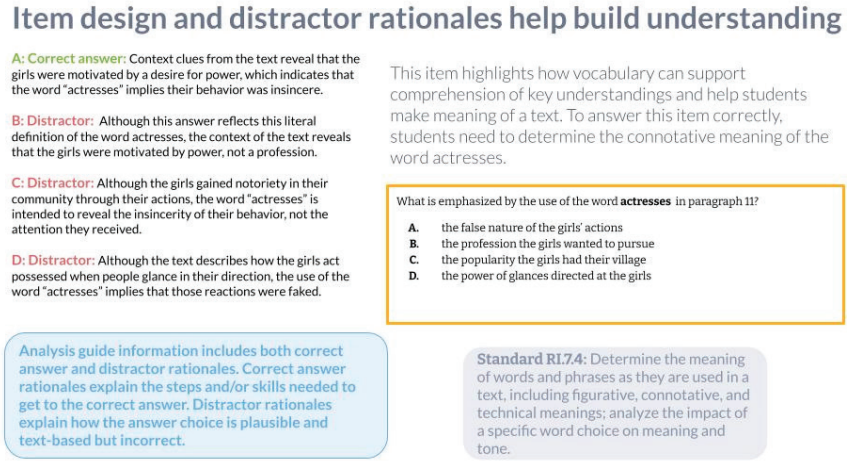
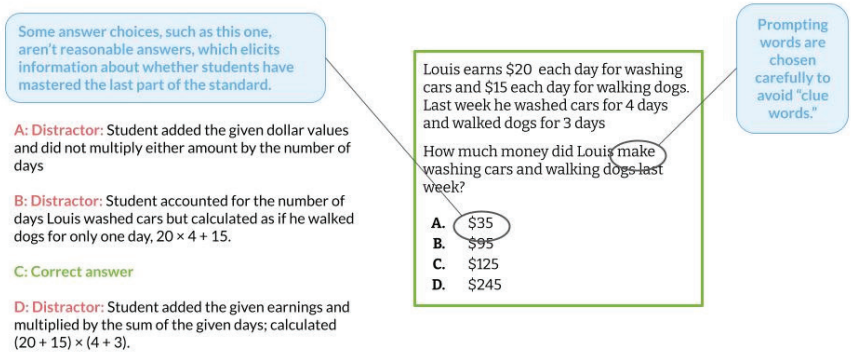


Figure 4.
ANet Distractor Rationale Math



Empowering Educators with Clear, Actionable Reports

For assessment transparency to meaningfully support instructional decision-making, results must be clear, timely, and directly applicable. Transparency in assessment design is only one piece of the puzzle. Equally important is ensuring that educators can easily interpret and act on assessment results through clear, structured reporting. Without structured, educator-friendly reporting, even the most well-designed assessments risk becoming stagnant data.

Research underscores that transparent reporting enhances instructional coherence by making assessment insights actionable across both classroom and system levels (Marion, Pellegrino, & Berman, 2024). When assessment results explicitly reveal student reasoning processes, teachers can identify misconceptions, interpret patterns in student thinking, and adjust instruction accordingly. Transparent reporting that includes descriptive feedback and contingent actions supports teachers in making real-time instructional decisions, enabling them to address learning gaps as they emerge rather than waiting for summative results. As Marion et al. (2024) emphasize, formative assessment must be an ongoing process, providing teachers with structured evidence to guide instructional adjustments, deepen student engagement, and support self-directed learning.

To operationalize these principles, ANet's reporting system is designed with usability in mind, ensuring that teachers can quickly and effectively interpret

assessment insights to inform instruction. Administered online, ANet's assessments provide fast, data-rich feedback, allowing educators to act on insights without delay. Rather than providing generic performance summaries, reports offer detailed, item-level analysis that helps educators pinpoint student reasoning, identify misconceptions, and adjust instruction accordingly. By emphasizing assessment as an ongoing instructional tool rather than a post-instructional evaluation, ANet's reports support responsive teaching and help educators take timely, targeted action to close learning gaps. ANet's reporting system is designed to be intuitive and meaningful, ensuring that teachers can:

- **Identify Specific Misconceptions:** Teachers can pinpoint patterns in student errors to understand not just what students got wrong, but why. This insight supports more targeted reteaching and intervention strategies.
- **Filter Student Performance by Key Indicators:** Reports allow teachers to sort and analyze student responses by standard, question type, and response patterns, enabling precise instructional targeting. This granular visibility helps educators adjust lessons in real time, ensuring that every student's needs are met.
- **Integrate Qualitative and Quantitative Data:** A combination of numerical scores, item rationales, and student work analysis provides a holistic view of student understanding, helping educators make data-driven decisions with confidence.

Research underscores the importance of presenting assessment results in a way that fosters instructional dialogue. Pellegrino et al. (2001) emphasize that assessment data should clearly define learning expectations and provide explicit criteria for student success. By including explanations of student work, reports enable teachers to engage in meaningful discussions about student progress and instructional next steps.

Collaborative Data Analysis Through Professional Learning Communities

Rather than reviewing data in isolation, educators can use ANet reports to facilitate professional learning conversations, leveraging reporting capabilities that enable item-level analysis, data disaggregation, and the creation of custom student groups. These tools allow teachers to facilitate professional learning conversations, discussing:

- What misconceptions are most common across student groups?
- Which instructional strategies have been effective, and where do adjustments need to be made?
- How can educators ensure that assessment insights translate into immediate instructional action?

> *"The data reports from ANet help us to target student strengths and weaknesses. Also, the assessments…give teachers and leaders examples of what students should know and be able to do if they have mastered a standard."*
> —Instructional Coach, Massachusetts

For many districts, the ability to analyze assessment data in collaborative PLCs has strengthened instructional coherence and decision-making. Madison Metropolitan School District (MMSD) serves as a powerful example of how transparency in assessment reporting can drive system-wide instructional improvement.

Case Study: Madison Metropolitan School District—Building a System of Data Transparency and Instructional Alignment

The Madison Metropolitan School District in Wisconsin serves over 27,000 students across 52 schools. Despite a strong commitment to improving student outcomes, the district faced significant challenges with assessment strategy and data use.

The Challenge: Fragmented Assessment Systems and Unclear Data Use

Before partnering with ANet in 2018–2019, MMSD lacked a cohesive and transparent approach to assessment. Teachers administered multiple assessments, yet the data provided little instructional value, making it difficult for them to adjust their teaching effectively. Additionally, district departments operated in silos, causing misalignment between assessment practices, curriculum goals, and instructional priorities.

District leaders recognized that to improve instruction, they needed a transparent and aligned assessment system that provided:

- Clear and accessible data that teachers could interpret and apply in real time
- A shift from data collection for accountability to data that actively informed instruction
- Stronger collaboration across departments to ensure a unified approach to assessment and instructional support.

The Solution: Creating a Clear and Actionable Assessment Strategy
MMSD took a multi-year, strategic approach to restructuring its assessment practices. With ANet's support, the district launched a comprehensive assessment strategy focused on transparency, alignment, and usability.

One of the district's first steps was conducting a district-wide assessment audit. The Assessment Priority Project working group, led by Caroline Racine Gilles, Executive Director of Integrated Supports and Assessment for Learning, evaluated every assessment in use across grade levels. Through this process, they identified redundant, misaligned assessments and prioritized those that best supported instructional decision-making.

> *"We recognized that we had a glut of evaluative assessments, which indicated the need to incorporate assessments closer to instruction. We want data to inform instruction, and we want to use data—both qualitative and quantitative—to engage students and families."—Caroline Racine Gilles, MMSD*

MMSD also standardized data reporting structures to ensure teachers could analyze student work, track performance trends, and make informed instructional choices. Instead of receiving broad performance summaries, educators were provided with detailed, item-level analysis that helped them understand not just what students got wrong, but why.

Additionally, the district embedded data conversations into PLCs. Rather than treating assessment data as a one-time event, teachers engaged in ongoing discussions about how to apply insights to daily instruction.

The Impact: Strengthening Educator Confidence in Data Use
The district's focus on data transparency and instructional alignment led to measurable improvements in educator confidence and instructional clarity.

- **93% of school leaders** agreed that the district had clearly stated instructional priorities, up from 56% before the initiative.
- **97% of school leaders** supported the district's vision for how assessments should be used in their schools, compared to only 44% previously.
- **100% of school leaders** supported the district's vision for using data in decision-making, an increase from 67% before the strategy was implemented

This shift meant that teachers spent less time assessing for compliance and more time using data to inform instruction. Educators now had the tools to make "just in time" adjustments to their teaching, ensuring that students received the support they needed when they needed it.

Lessons Learned: Key Takeaways from MMSD's Data Transformation
MMSD's experience underscores several essential lessons for districts looking to strengthen data transparency and instructional alignment:

- Clear assessment data enables teachers to make instructional decisions with confidence.
- Assessment must be embedded within professional learning structures so that data is not just collected but actively used.
- Coherence between assessment, curriculum, and instructional goals is critical to ensuring that data serves as a tool for learning rather than compliance.

Looking Ahead: Connecting to Coherence
MMSD's commitment to assessment transparency improved data use and laid the foundation for deeper instructional coherence. By ensuring that assessment, curriculum, and instruction were aligned and mutually reinforcing, the district moved beyond transparency to create a more unified and effective educational system.

Sustaining this level of coherence requires more than visibility into assessments—it demands an intentional focus on instructional time, strategic assessment design, and alignment with high-quality instructional materials (HQIM). This balance between clarity and coherence is key to ensuring that assessments serve learning rather than disrupt it.

Designing for Coherence: Reducing Testing Overload and Strengthening Instruction

One of the most powerful lessons from our work with districts has been that transparency is only the beginning. Initially, providing educators with clear, accessible assessment resources and design structures helped schools understand what students would be tested on and why. This approach ensured that assessments were not a mystery but a meaningful part of the instructional process.

However, transparency alone does not guarantee coherence. True instructional coherence depends on more than assessment alignment—it requires a learning climate where instruction, not excessive testing, is the focus. When schools are oversaturated with assessments, instructional time becomes disrupted, leaving little room for deep learning experiences. Teachers struggle to find meaningful takeaways from overwhelming amounts of data, and students experience assessments as interruptions rather than opportunities for growth.

To build a stronger learning environment, districts must first reduce the number of assessments that compete for instructional time. By streamlining assessment systems, ANet helps districts shift their focus from excessive measurement to actionable insights, ensuring that assessments serve learning rather than disrupt it. As Caroline Racine Gilles of MMSD observed, many districts face a *glut of assessments*, which creates unnecessary testing burdens without providing meaningful instructional value. This effort of ensuring that data informs teaching begins by tackling the Volume Problem.

The Volume Problem: Fewer, Better Assessments

In K–12 education, assessments are designed to serve many purposes, from enhancing learning to ensuring accountability. While they aim to cover various educational needs—diagnostic, formative, summative—the sheer proliferation of assessments has led to an unintended consequence: over-testing that drains instructional time while offering little actionable insight.

Despite increased spending on assessments (Simba Information, 2019), the anticipated improvements in educational outcomes have not materialized. The Council of Great City Schools (Hart et al., 2015) reported that K–12 students spend an average of 20 to 25 hours per year taking standardized tests—a figure that does not account for test preparation time, which can push the total to over 100 hours annually when including interim and locally developed assessments.

Additionally, while most educators are not data scientists, they rely on assessment data to inform instruction. However, the volume and variety of assessments, ranging from interim to high-stakes summative testing, creates a chaotic landscape where teachers sift through excessive data that often lacks coherence and alignment.

As a result, instead of supporting student learning, assessments risk becoming obstacles, consuming valuable class time without always providing meaningful insights that drive instructional improvement.

Streamlining Assessment: How ANet is Reducing Testing Time While Enhancing Insights

ANet's early assessments were designed to cover the full depth and breadth of academic standards, ensuring alignment with rigorous instructional goals. However, as schools implemented these assessments, a clear challenge emerged: ensuring that assessment length remained practical in an already oversaturated testing environment. Many sessions took longer than a single class period, disrupting instruction rather than supporting it. This raised a fundamental question: How can we maintain rigor while reducing assessment time?

Recognizing the need to balance depth with efficiency, ANet began developing streamlined assessments—shorter in length but equally rigorous and instructionally meaningful. The goal was to reduce testing time while maintaining instructional

value, ensuring that assessments were a tool for learning rather than an interruption. While these improvements made assessments more manageable, another challenge remained: teachers needed time and capacity to analyze student results and make instructional decisions.

Even though ANet provided teachers with insights into assessment design and common student misconceptions, acting on this information required time—an increasingly scarce resource for educators. This led to a new phase of experimentation: Could assessments be further optimized to provide the same depth of insight in significantly less time? ANet is currently exploring ways to leverage adaptive assessment models and predictive insights to create assessments that are not only shorter but also more precise in identifying student needs. While early efforts show promise, this work remains ongoing, with a focus on ensuring that any reductions in assessment length enhance instructional value rather than diminish it.

Reimagining Math Assessments

For nearly two decades, ANet has been collecting and analyzing student misconception data, allowing for a deeper understanding of the predictable errors that hinder math proficiency. Through this extensive dataset, ANet can now anticipate where students are likely to struggle, shifting the role of assessment from a reactive tool for remediation to a proactive tool for prevention.

ANet is redesigning math assessments to provide more actionable insights in less time. Traditional math assessments often focus solely on correctness, missing the opportunity to understand *why* students make errors. By leveraging misconception trends alongside adaptive diagnostics and machine learning technology, ANet's next generation of math assessments aims to:

- Predict when and why students will struggle through adaptive diagnostic assessments, allowing teachers to plan student-level interventions and system-wide supports.
- Track student growth and mastery in real-time with short, monthly assessments that are standards-based and formative, helping educators measure how well students internalize instruction.

- Take significantly less time than traditional testing, freeing up instructional hours while still delivering deep insights.
- Provide students with immediate feedback on their strengths and areas for growth, fostering confidence and engagement.

By integrating machine learning, adaptive diagnostics, and real-time progress tracking, ANet's math assessments will provide a clearer, more efficient picture of student learning. Teachers spend less time testing and more time teaching, while still getting the insights they need to drive instruction forward.

As we continue to refine assessment practices, a key question remains: How can we innovate while preserving the quality and depth that make assessments valuable? Reducing testing time is an important step, but true instructional impact depends on more than efficiency.

Even when assessments are well-timed and structured for ease of use, their true impact depends on deeper factors—whether they accurately capture student thinking, generate meaningful insights, and align with instructional goals. Without these elements, assessments serve as compliance exercises rather than powerful levers for student success.

To be truly effective, assessments must do more than exist within structured timelines; they must be designed with precision—asking the right questions, uncovering student reasoning, and guiding instructional decisions. Yet, too often, assessments fall short. Gaps in alignment, ineffective item design, and static reporting structures weaken their value, leaving educators without the insights they need to foster student growth. Addressing these gaps requires a fundamental shift—one that ensures assessments are not just efficient, but instructionally powerful, fully aligned to curriculum, and responsive to the way students learn.

The Quality Problem: Strengthening Alignment for Meaningful Assessment

The quality of assessments in education is a layered challenge, impacting how well they enhance learning rather than just measure it. Research shows that assessments often fail to align with curriculum, capture student thinking, or provide actionable insights for educators. These issues fall into three interrelated dimensions:

Misalignment Between Assessments and Learning Goals

Despite increasing adoption of HQIM, many assessments remain disconnected from curriculum limiting their instructional value. Pellegrino et al. (2001) argue that valid assessments must align with both curriculum standards and cognitive learning processes. Davidson, Shepard, & Penuel (2017) further highlight the need for coherence across curriculum, instruction, and assessment to avoid superficial test-based learning. Without such alignment, assessments become isolated measures rather than integral learning tools.

Weak Item Design Fails to Surface Student Thinking

Traditional items often fail to diagnose student misconceptions, reducing assessment to a binary right/wrong judgment. Poorly designed questions provide little insight into how students reason through problems, making it difficult for teachers to adjust instruction effectively.

Black and Wiliam (1998) demonstrate that assessments must not only classify student performance but also reveal underlying thinking to inform targeted interventions. Similarly, Heritage (2010) emphasizes the importance of designing assessments that allow educators to identify misconceptions in real time. Without this diagnostic function, assessments risk reinforcing rather than addressing learning gaps.

Inadequate Reporting Limits Instructional Utility

Even well-constructed assessments lose impact if their results do not provide clear, actionable insights for educators. Traditional reports often emphasize scores over meaningful data, limiting teachers' ability to adjust instruction.

Elmore (2019) critiques large-scale assessments for prioritizing ranking over learning, arguing that assessment systems should provide feedback that enables educators and students to take informed action. The National Research Council

(2011) similarly advocates for reporting that translates data into instructional guidance, rather than frozen performance snapshots. Without accessible, transparent reporting, assessment data remains underutilized.

These challenges illustrate that assessment quality is not only about test design. It is a systemic issue. Effective assessments must be coherent with curriculum, diagnose misconceptions, and generate actionable insights. For assessments to move beyond isolated measures of learning, they must be designed within a system of coherence, where curriculum, instruction, and assessment are seamlessly integrated. Without this alignment, even well-designed assessments risk being misused or disconnected from the learning process. Ensuring that assessments work in tandem with instruction requires a system where curriculum, teaching, and assessment reinforce one another. This level of coherence is essential for creating structured, equitable, and effective learning experiences.

Coherence: Building Alignment Across Curriculum, Instruction, and Assessment

Coherence is achieved when assessments, instructional practices, and curriculum are seamlessly aligned, ensuring that students receive a structured, equitable, and effective learning experience. Research supports this integrated approach, as Marion et al. (2024) emphasize that balanced assessment systems—those that integrate formative assessments with instructional practices—are essential for ensuring all students receive the support needed to achieve excellence. When coherence is present, students progress through a thoughtfully designed system where each stage of learning builds on the previous one, guided by clear expectations and meaningful assessments. However, coherence is often disrupted when curriculum adoption, professional learning, and assessments operate in silos, resulting in incoherent instructional practices that fail to support student learning effectively.

Theoretical Framework for Coherence

A well-designed assessment system does not operate independently of instruction but rather serves as a reinforcing mechanism within a broader instructional model. Pellegrino et al. (2001) argue that the model of learning should serve as a unifying element that brings cohesion to curriculum, instruction, and assessment. Without this cohesion, assessments risk measuring knowledge in isolation, providing data that lacks instructional relevance. When assessments are not synchronized with instruction and curriculum, the learning process becomes fragmented. Pellegrino

et al. (2001) note that if any of these components are misaligned, the balance of the system is disrupted, leading to misleading assessment results or ineffective instruction. Achieving coherence requires thoughtful coordination to ensure that assessments not only measure learning but actively support it.

Educational coherence is further complicated by the fact that curriculum, instruction, and assessment operate at multiple levels. State policies may set assessment requirements, districts make curriculum choices, and teachers determine instructional methods. Pellegrino et al. (2001) emphasize that these layers of decision-making require ongoing adjustments to maintain coherence, both horizontally within districts and vertically across state, district, and classroom levels.

Recognizing the necessity of coherence, ANet's approach to assessment design has evolved over time. Initially, assessments were structured around standards alignment, with the expectation that schools would bridge the gap between curriculum and assessment. However, as districts began adopting better instructional resources to guide instruction, it became clear that assessment design needed to reflect these curricular structures more intentionally. Yet, simply aligning assessments with HQIM is not enough. Coherence also depends on how well teachers are prepared to implement these materials in practice.

Challenges of HQIM Implementation and the Role of Assessment

Quality instructional materials are a critical lever for improving student outcomes, yet research shows that teachers frequently supplement or replace district-adopted HQIM with resources of uncertain rigor or alignment (Steiner, 2024). Steiner argues that this behavior 'ensures that the material a child studies in school differs from classroom to classroom' and that 'the caliber, rigor, and any rational sequencing of that material both within and across grade levels becomes a matter of luck and chance.' Given these inconsistencies, assessments serve not only to measure student proficiency but also to verify whether HQIM is being used as intended, supporting instructional alignment across classrooms.

Achieving coherence requires more than alignment as an idea. It thrives when teachers have the knowledge, resources, and support to bring materials to life in the classroom. Professional learning ensures that teachers can effectively implement HQIM, interpret assessment data, and make informed instructional adjustments that keep student learning on track. By embedding professional learning into the

instructional cycle, ANet strengthens the connection between assessments and teaching, helping educators translate insights into action.

Building Proactive Professional Learning

Rather than waiting until students fail assessments to recognize misconceptions, ANet is currently exploring how predictive insights can be integrated into professional learning, testing ways to equip educators with the tools to prevent misunderstandings before they occur. A key part of this approach is ensuring that assessments and instructional decisions are anchored in HQIM, reducing the reliance on inconsistent supplemental resources. By aligning professional learning with HQIM, ANet helps educators maximize the effectiveness of their curriculum, reinforcing instructional coherence across classrooms and grade levels. This process includes:

- Analyzing historical assessment data to pinpoint common misconceptions at each grade level.
- Preparing teachers with targeted professional learning before content is taught, equipping them with strategies to address predictable challenges aligned to current curriculum sequencing.
- Post-assessment reflection, where educators analyze student performance, assess instructional adjustments, and refine teaching strategies.
- Ongoing refinement through teacher feedback, ensuring continuous improvement of instructional approaches.

By embedding misconception-driven Professional Learning (PL) into the teaching cycle, ANet hypothesizes that:

- Teachers who receive PL on guided adaptations of HQIM will make more meaningful adjustments that enhance learning opportunities.
- Students whose teachers implement these guided adaptations will perform better on the targeted math content.
- Teachers will develop a stronger perception of HQIM quality and usability, leading to more effective curriculum implementation.

By combining predictive professional learning with redesigned assessments, ANet is positioning assessments not just as reflections of past learning but as guides for future instruction. This integrated model ensures that students receive the right

support at the right time—before misconceptions take hold—helping them build a stronger foundation for long-term success.

Achieving this level of coherence is not simply a matter of aligning assessments with HQIM. It also requires ensuring that teachers are equipped to implement HQIM with fidelity. When assessment data reveals gaps between intended and actual implementation, it signals where professional learning can provide targeted support, reinforcing HQIM rather than replacing it. This approach has been central to district-level successes, such as Carlsbad Municipal Schools, where the alignment of assessments, curriculum, and professional learning created a more coherent instructional system.

Coherence in Action: The Carlsbad Case Study

The Challenge of Inconsistency

Before undertaking its instructional transformation, Carlsbad Municipal Schools struggled with instructional inconsistency and low curriculum fidelity. While HQIM had been adopted, teachers often supplemented the curriculum with external resources, leading to significant variation in instructional pacing and rigor across schools. Without clear alignment, assessments were unable to accurately measure instructional effectiveness, reinforcing inequities rather than addressing them.

A Systemic Approach to Coherence

Carlsbad's leadership recognized that coherence required more than just aligning assessments to HQIM. It required a system-wide shift in instructional priorities. With ANet's support, the district:

- Established a transparent curriculum selection process that engaged teachers and leaders in decision-making, building trust and buy-in.
- Used assessment data and instructional observations to identify where HQIM was not being implemented with fidelity.
- Created an instructional leadership department focused explicitly on coherence, professional learning, and curriculum implementation.

The Role of Assessment in Verifying Implementation
To ensure fidelity, Carlsbad used a combination of:

- Formative assessments aligned to HQIM to track student progress.
- Instructional walkthroughs to observe whether teachers were delivering grade-level content as intended.
- Targeted professional development informed by assessment data to help teachers adjust instruction while maintaining curriculum integrity.

As a result, instruction became more consistent across schools for the first time. Principal Stacy Rush noted, "*You could go into several Algebra I classrooms, for example, and you would see they were in the same place. Finally, we had coherence and consistency.*"

Sustaining Coherence Through Leadership
To make these changes lasting, Carlsbad established a district-wide instructional leadership team dedicated to supporting strong, standards-aligned instruction. Leaders participated in professional learning and used assessment data strategically to refine instructional approaches, ensuring that coherence was not just a one-time initiative but an ongoing priority.

Toward a Fully Coherent System
A coherent instructional system ensures that assessments are not separate from, but rather embedded within, the learning process. District and school leaders must work together to create an infrastructure where:

- HQIM is implemented with fidelity through aligned professional development.
- Assessments are streamlined and transparent, providing actionable insights for teachers.
- Instructional leadership prioritizes coherence, ensuring that teachers are not left to navigate curriculum and assessment misalignment on their own.

Carlsbad's transformation underscores a critical takeaway: coherence is not simply about aligning curriculum and assessment on paper. It requires intentional leadership, professional learning, and assessment-informed instructional adjustments. By committing to a more coherent system—where district initiatives,

instructional leadership, curriculum implementation, and assessment strategies reinforce one another—schools can create an environment where students receive the high-quality instruction they deserve, and educators are empowered to drive meaningful learning outcomes.

Student-Centered Design: Elevating Student Experience and Self-Efficacy

Ensuring all students have access to high-quality learning experiences requires a transformation in how assessments are designed and used. Many assessments prioritize prediction over intervention, lack transparency in item design, and can fail to disaggregate data in ways that allow for targeted instructional support (Davidson, Shepard, & Penuel, 2017). These limitations disproportionately impact students who require differentiated learning pathways, making it difficult for educators to address specific needs effectively.

Too often for students who have the hardest time in traditional classroom structures, assessments are used as tools weaponized against them instead of empowering them. Yet, Black and Wiliam (1998) demonstrated that formative assessments—when integrated into instruction—improve student learning outcomes, particularly for historically under-served students. When students engage with assessment as a reflective practice rather than a judgment, they gain agency over their learning, which Elmore (2019) argues is essential for fostering deeper cognitive development. This shift turns assessments from obstacles into pathways for student success. When designed to highlight strengths and guide learning, assessments strengthen student agency, self-efficacy, and a path to growth and achievement.

Assessment must be responsive to students' lived experiences. Ladson-Billings (1995) introduces culturally relevant pedagogy, stating that equitable assessment must affirm students' identities while supporting academic achievement, and when assessments reflect the cultural backgrounds and experiences of students, they are more likely to engage and motivate learners, leading to improved academic performance. Similarly, Paris and Alim (2017) advocate for instructional systems that recognize students' diverse backgrounds, emphasizing that assessments should sustain students' identities rather than impose deficit-based frameworks. To truly serve all students, assessments must reflect the rich diversity of their experiences, languages, and ways of knowing.

Often, traditional assessments lag behind curriculum advancements and fail to provide culturally and linguistically inclusive representations, limiting engagement and missing opportunities for deeper learning. When assessments incorporate culturally relevant content and allow multiple ways for students to demonstrate understanding, they foster deeper engagement and more accurate measures of learning. By designing assessments in this way, we move beyond exclusionary models toward systems that validate, challenge, and support every learner's success.

To achieve truly equitable, student-centered design, assessment must shift from a tool of evaluation to an opportunity for meaningful engagement. This evolution is key to fostering instructional coherence, strengthening leadership accountability, and building transformative school cultures because at its core, assessment must serve and engage students. Achieving this requires assessment design that prioritizes engagement, transparency, and student agency.

Shifting from Evaluation to Engagement

Assessments should serve learning, yet students often experience them as isolated, high-stakes events. Without transparency into why certain content is assessed or how results shape instruction, assessments feel disconnected from the learning process.

When assessment design is transparent, students engage more deeply. They see purpose and relevance in the content, making assessments a continuation of their learning experience rather than a separate, evaluative task. They also develop a greater sense of agency over their learning, as they understand what is being asked of them and why. By designing assessments with transparency in mind, ANet ensures that both educators and students receive actionable insights that drive learning rather than prediction.

Connecting the Student Experience

To ensure students engage meaningfully with assessments, ANet integrates relevant themes and real-world connections into its content. This approach strengthens the link between learning and assessment, increasing motivation and deepening understanding. By embedding themes that reflect diverse student experiences, ANet ensures that assessment tasks are rigorous, relevant, and

connected to students' identities. To support this alignment, ANet designs assessment content to reflect high-quality curriculum standards, reinforcing connections between classroom instruction and assessment outcomes (EdReports, 2021).

In alignment with *EdReports' Gateway 3* - Usability Criteria, assessments are designed to measure student progress and to promote meaningful engagement through texts that reflect diverse perspectives and experiences (Criterion 3.3). By ensuring that assessments are both rigorous and reflective of classroom instructional materials, students can better connect their learning to their assessments, strengthening engagement. As a result, ANet aligns content to high-quality curriculum standards to reinforce coherence between what students learn in class and what they are assessed on. To foster engagement and accessibility, ANet ensures that assessments feature a diverse range of voices, historical perspectives, and meaningful themes. Nearly half of ELA passages feature female protagonists or historical figures, and a majority highlight individuals from a variety of cultural backgrounds, ensuring that students see themselves—and others—reflected in what they read through the lens of both diverse achievements and everyday life experiences.

When teachers have full visibility into assessment design, they can explain its purpose to students and ensure assessments align with what students have been learning, reducing disconnects in engagement. Black & Wiliam (1998) show that formative assessments integrated into instruction improve student outcomes, particularly for historically underserved students. Students engage more deeply when they see connections between what they learn and what they are assessed on. When assessments reflect diverse perspectives while maintaining rigorous academic standards, all students feel included in the learning experience.

Assessment as Transparent Dialogue: Looking at Student Work & Misconceptions

Engagement is about more than interaction. It's about ownership. When students understand their own progress, they can set goals, self-reflect, and take an active role in their learning. Transparent assessment reporting transforms assessments from isolated evaluations into dynamic feedback loops that support student agency. Assessment should not be a one-way process where students complete a test and simply receive a score. Engagement doesn't stop at taking an

assessment—it must extend into the reporting process. Often, students are given results but no insight into their thought processes, misconceptions, or how their responses connect to future learning. This approach misses a critical opportunity for engagement—one where students learn to analyze their own reasoning and develop greater self-efficacy.

Transparent reporting encourages students to interrogate their own choices: to reflect on their learning, identify patterns in their thinking, and refine their approaches to problem-solving, reinforcing a growth mindset. When students reflect on why they answered a question a certain way, they strengthen self-efficacy and build the metacognitive skills necessary for long-term learning (Elmore, 2019).

Traditional assessment reporting usually focuses on correctness, not on understanding, leaving students without clear next steps. ANet's reporting system provides teachers and students with insight into student reasoning by highlighting misconceptions embedded in incorrect responses so students can reflect on their thinking. The goal is to encourage discussions around student work both within PLCs and directly with students, allowing students to articulate their reasoning and learn from their mistakes, offering real-time insights that connect assessment outcomes to targeted instructional strategies and future learning.

Teachers facilitate classroom discussions that prompt students to ask,"*What was my reasoning for choosing this answer?*" When students actively engage with their results, they take ownership of their learning, recognizing patterns in their mistakes and helping them make adjustments in real time. They become participants rather than passive recipients of assessment outcomes. Students begin to see mistakes as part of their learning process rather than as indicators of failure.

From Judgment to Conversation

Creating an equitable assessment system requires shifting the conversation—from using assessments as final judgments of ability to positioning them as opportunities for growth and reflection. Equitable assessment does not mean lowering expectations—it means ensuring that students understand what is being asked of them, why it matters, and how they can grow from the experience. Student-centered design relies on transparency and coherence. It is essential because it enables students to engage more deeply when they can see themselves

in assessment content. Teachers can better support students when they have full insight into how assessments align with instruction.

Assessment must move beyond evaluation. It must engage, inform, and empower. By prioritizing transparency and student-centered design aligned to quality materials, we transform assessments from a system of judgment into a tool for continuous learning and growth

Student Agency: Case Study on Foundational Literacy and Reading Confidence

Like most traditional assessments, ANet's assessments are designed to illuminate comprehension mastery, ensuring that students can analyze and engage with complex texts. However, emerging data and research highlight a critical gap: students who struggle with foundational literacy skills—decoding and fluency—may be unable to fully access comprehension-based assessments. Before these students can analyze what they read, they need support in building the skills and confidence necessary for reading engagement.

National trends underscore the urgency of this issue. National Assessment of Educational Progress (NAEP) data indicates that over 60% of 4th graders, 8th graders, and 12th graders are reading below the proficiency level, meaning they have not yet reached grade-level reading expectations (National Center for Education Statistics, 2023). The reality is stark—many older students have not yet developed the foundational skills needed to support comprehension. Without intervention, these challenges compound over time, diminishing students' motivation to engage with reading altogether.

The Intersection of Reading Proficiency and Reading Identity

Studies show that students who struggle with reading for an extended period begin to internalize a negative reading identity, seeing themselves as non-readers (Learned, Frankel, & Brooks, 2022). The longer students face difficulty with decoding and fluency, the less likely they are to identify as readers, engage with literacy-based tasks, or seek out opportunities to practice. This lack of engagement leads to fewer reading experiences, which in turn makes improving literacy skills even more difficult—a phenomenon known as the "Matthew Effect" in reading development (Stanovich, 1986).

In middle and high school classrooms, this struggle manifests in subtle yet significant ways: students who lack confidence in reading often avoid participation, experience anxiety when asked to read aloud, and disengage from texts that appear too challenging. This loss of reading agency not only affects academic outcomes but also deepens educational inequities, as students with weaker foundational skills are left further behind.

A New Approach: Pairing Foundational Literacy Assessment with Student Confidence Measures

To address this growing crisis, ANet partnered with Reading Reimagined, supported by AERDF and Stanford University, to launch a pilot program in district middle and high schools. This initiative featured the ROAR assessment—Rapid Online Assessment of Reading—developed at Stanford University, a groundbreaking tool designed to evaluate foundational reading skills in students from grades K–12.

The ROAR assessment provides a comprehensive, gamified online experience, measuring key foundational literacy skills, including:

- Phonemic awareness
- Word-level decoding
- Sentence-reading fluency

The fully online process takes 30 minutes or less, offering quick yet invaluable insights into students' foundational reading abilities. This allows educators to pinpoint gaps in decoding and fluency that might otherwise go unnoticed, particularly among older students who are expected to engage in comprehension-based assessments without adequate foundational support.

At the same time, the **Motivation to Read Profile**, rooted in research from Gambrell (1996), measured students' self-efficacy in reading, providing teachers with critical insight into how students feel about reading, including their confidence in reading aloud and their overall attitude toward literacy tasks.

The Impact of Reading Confidence on Literacy Development

Early results from the pilot revealed a significant lack of confidence among struggling readers, with many students expressing deep anxiety about reading in front of peers. One student candidly shared:

> *"I want my teachers to know that I sometimes [struggle] when I read out loud in class. I get stuck on a word that is hard for me to pronounce. And sometimes I pronounce words wrong, which can be difficult. So, thanks for your understanding."*

This emotional barrier is a key factor in literacy development. Students who lack confidence in their reading ability often avoid engaging with texts, reinforcing the cycle of low literacy and low motivation. However, by pairing diagnostic literacy assessments with measures of self-efficacy, educators can better understand both the skill-based and psychological barriers to reading success.

ANet's work in foundational literacy is still evolving, but the early findings are clear: addressing decoding and fluency is just as important as assessing comprehension, particularly for older students who have struggled to develop strong literacy foundations. Our evolving approach to literacy assessment mirrors this shift, moving beyond evaluation toward assessments that directly support student confidence, engagement, and foundational literacy growth.

Advancing Student-Centered Design with Adaptive Learning Technologies

As ANet continues to refine its approach to assessment, one fundamental principle remains constant: assessments must not only be rigorous but also accessible. This means ensuring that all students—regardless of their starting point—can fully engage with grade-level content in ways that foster both mastery and growth.

A key part of this vision is deepening student-centered design, recognizing that students learn best when they feel confident in their abilities and see assessments as a tool for growth rather than a judgment of their abilities. Research has shown that student efficacy increases when they have a sense of agency in their learning. As a result, ANet is exploring ways to:

- Expand student choice in assessment formats, ensuring that students can demonstrate understanding in ways that reflect their strengths.
- Integrate culturally relevant content, making assessments more engaging and reflective of students' lived experiences.
- Balance mastery and growth, maintaining high academic standards while providing differentiated access to content based on student readiness.

With these priorities in mind, the future of assessment must embrace adaptive learning technologies. By integrating AI-driven insights and adaptive assessment models, ANet is working toward a system in which assessments dynamically adjust to student responses, ensuring that every student is met at the right level. These adaptive assessments combine diagnostic and mastery-based items, reducing test-taking time while simultaneously improving the precision of insights for teachers.

Advancements in machine learning and real-time data analytics also open new possibilities. Technology allows for a sharper focus on diagnostic purposes, helping educators pinpoint not just where students struggle, but why. As these innovations take hold, ANet continues to ground its work in the lessons learned over the past twenty years. Understanding what makes assessments truly effective and where traditional approaches fall short has shaped the evolution of our design. These insights guide our commitment to ensuring that assessments actively support student growth rather than serving as rigid measures of ability.

Results & Impact: Lessons That Shape Assessment Design

For twenty years, ANet has worked alongside schools and districts to transform how assessment fuels instruction. Our approach—pairing high-quality, instructionally focused formative assessments with targeted professional learning—has helped educators make better, data-driven decisions for student success. However, our journey has also revealed critical insights about what makes assessments truly effective and where traditional approaches must evolve.

A key question in a 2010 federal innovation grant (i3), analyzed by Harvard University, was whether timely student performance data—paired with targeted support—could improve instructional practices and boost student achievement. The answer? It depends. While the study showed that ANet's program improved

teacher data usage and instructional decision-making, student achievement gains were only significant when schools had the right readiness conditions in place (West, Morton, & Herlihy, 2016). In short, data alone wasn't enough. Impact depended on whether teachers and schools had the capacity to act on it.

This was a critical insight: Great assessments alone aren't enough. Their impact depends on whether teachers and schools are ready to act on the data. In response, ANet recalibrated its approach, expanding its focus beyond school-level data literacy to ensure that assessment-driven success is supported at every level of the system. Our adaptive strategies now strengthen vertical coherence from the district office to the classroom, enhancing the implementation of high-quality curricula and developing assessments that provide timely, actionable insights to improve teaching and learning.

Demonstrating Efficacy Through Continued Evaluation

While ANet is committed to designing student-centered and adaptive assessments, the ultimate measure of success is whether these assessments lead to improved outcomes. To ensure that our innovations are effective, ANet employs a rigorous, data-driven evaluation process to assess the impact of our work.

Internal evaluations consistently demonstrate that ANet-supported schools show stronger performance on summative assessments than comparable non-ANet schools. However, a simple comparison does not fully capture the depth of our impact. To isolate ANet's direct effect on student learning, we use a three-step evaluation process:

1. **Matching:** Each ANet partner school is paired with a non-ANet school of similar demographics and prior achievement levels.
2. **Change Calculation:** We track how performance changes over time in ANet-supported schools versus their matched counterparts.
3. **Difference-in-Difference Analysis:** The differential in performance growth between ANet and non-ANet schools allows us to quantify ANet's direct impact.

This method ensures that our results are not just anecdotal but backed by empirical evidence. We consistently observe that when readiness conditions are in place, ANet's coaching, assessments, and instructional strategies lead to measurable

improvements in student learning. The takeaway is clear: assessments only drive improvement when they exist within a system that supports teachers in acting on them. Readiness conditions, targeted coaching, and aligned instructional practices determine whether assessment leads to meaningful change. Schools like Honey Dew Elementary put these insights into action, leveraging data, refining instruction, and demonstrating what is possible when assessment is used as a tool for learning rather than measurement.

Case Study: Transforming Educational Outcomes at Honey Dew Elementary

The story of Honey Dew Elementary School in Renton, WA, exemplifies how a strategic approach to assessment, professional learning, and responsive instruction can transform student outcomes. Over the course of their ANet partnership, Honey Dew saw a 12.2% positive change in math proficiency compared to their matched comparison group of non-ANet partners. Their journey offers a powerful case study in how schools can move beyond a strong culture to drive measurable academic success.

When Principal Misty Mbadugha joined Honey Dew in 2014, she inherited a school with a positive culture but a lack of academic rigor. Recognizing the need for change, she sought to elevate instructional quality and ensure assessments were used as tools for learning rather than just measurement. In 2019, Honey Dew partnered with ANet to integrate a structured teaching and learning cycle—one that would align assessments with instruction and professional development.

Strategic Implementation: Turning Data into Action

From the outset, the school's leadership team, including Title I Math Coach, Becca L'Amour, and ELA Instructional Facilitator, Brooke Argotsinger, worked closely with ANet coaches to refine their approach to data-driven instruction. This partnership focused on helping teachers not only understand assessment results but use them to inform targeted interventions.

The turning point came when a professional learning session did not go as planned, prompting the team to rethink their instructional approach. This led to a shift toward more interactive and reflective professional development, helping teachers use ANet assessments to diagnose and address specific student needs in real time.

From Assessment to Impact: The Story of David

One of the most vivid examples of this transformation is the story of David, a fifth-grader struggling with fractions. Through ANet's interim assessments, his teacher was able to pinpoint his specific challenges and provide targeted instruction that rebuilt his confidence in math.

David reflects on the role of these assessments in his learning:

> *"It's important for your teacher to know what you need to learn. If you rush through your test… then your teacher won't know what you need to work on."*

Rather than viewing assessments as a test to pass or fail, David saw them as an ongoing dialogue about his progress. He even acknowledged the value of making mistakes:

> *"When you're wrong, you always learn something from your mistakes."*

This shift in mindset—from seeing assessments as high-stakes evaluations to seeing them as learning tools—is central to ANet's vision. By using assessments to guide real-time instructional adjustments, Honey Dew created a culture where students and teachers alike were empowered by data.

Broader Impact and Continuous Growth

As teachers became more adept at using data, student engagement and academic performance improved significantly. By the 2020–2021 academic year, students at Honey Dew saw a 10% or greater improvement in 38% of the assessed standards year over year.

These gains were not just one-time improvements. They reflected a lasting shift in instructional leadership. Teachers were no longer just administering assessments. They were leveraging them as tools for responsive teaching.

The Takeaway: The Power of Readiness, Coaching, and Continuous Improvement

Honey Dew's transformation underscores a central theme of this chapter: Assessment is most powerful when it is embedded within a system that supports instructional leadership and continuous improvement.

The success at Honey Dew was not just about implementing assessments—it was about ensuring teachers had the professional learning, coaching, and leadership structures in place to use assessments effectively. This case study reinforces three key takeaways:

1. Assessment alone does not drive improvement—how educators use assessment data is what matters.

2. When readiness conditions are in place, ANet's coaching and instructional strategies lead to measurable and sustained student growth.

3. Continuous improvement is essential. Schools must be willing to adapt their strategies in response to both successes and challenges.

As ANet continues to refine its student-centered, adaptive, and data-driven assessment models, the lessons from Honey Dew serve as proof of concept for what is possible. Schools that invest in a structured teaching and learning cycle—one that integrates responsive assessment, professional learning, and strong instructional leadership—can achieve breakthrough results for students.

Honey Dew's journey exemplifies what is possible when assessment moves beyond a tool for accountability and becomes a driver of learning. Their success highlights the essential conditions for impact: a clear instructional vision, professional learning that enables teachers to refine their practice, and assessments that serve as formative tools rather than final judgments. This model not only transforms schools, it reshapes the role of assessment itself, proving that when assessment is embedded within a system of instructional coherence, real and lasting student growth follows.

Conclusion

Assessment has long been viewed as a necessary but imperfect tool, often associated with accountability rather than learning. But as schools rethink how assessments are designed and used, a different reality emerges: assessments can do more than measure learning; they can accelerate it.

Throughout this chapter, we have explored the fundamental shifts required to make assessments more transparent, coherent, and student-centered. We have seen that assessment systems must be embedded within instructional cycles, connected to high-quality curricula, and designed to provide meaningful, real-time insights that empower both teachers and students.

Schools like Madison Metropolitan School District (MMSD), Carlsbad Municipal Schools, and Honey Dew Elementary exemplify what is possible when assessment moves beyond passive evaluation and becomes an active driver of learning. MMSD strengthened transparency and instructional alignment. Carlsbad built coherence across curriculum, assessment, and professional learning, and Honey Dew leveraged assessments to transform instructional decision-making. Each of these schools demonstrates that when readiness conditions are in place, assessments can shift from being a source of compliance to a catalyst for meaningful student growth.

As ANet continues to refine its approach, we remain committed to the vision that assessments must not simply track progress, but actively contribute to it. The future of assessment is one in which data informs—not dictates—teaching and learning. And as schools embrace this future, they move closer to an educational system that truly puts students at the center of every decision.

References

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment.* King's College London, School of Education.

Davidson, K. L., Shepard, L. A., & Penuel, W. R. (2017). *Design principles for new systems of assessment. Phi Delta Kappan.*

EdReports. (2021). *ELA grades 3–8 review criteria v1.5.* https://www.edreports.org

Elmore, R. F. (2019). *The future of learning and the future of assessment. ECNU Review of Education, 2*(3), 328–341.

Gordon, E. W. (2013). *To assess, to teach, to learn: A vision for the future of assessment.* Learning for Action Research Network. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's Great City Schools: An inventory and preliminary analysis.* Council of the Great City Schools. https://files.eric.ed.gov/fulltext/ED569198.pdf

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Corwin Press.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491.

Learned, J., Frankel, K., & Brooks, M. (2022). Disrupting secondary reading intervention: A review of qualitative research and a call to action. *Journal of Adolescent Literacy,* (in press), 1-11.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

National Research Council. (2011). *Assessing 21st century skills: Summary of a workshop.* The National Academies Press. https://doi.org/10.17226/13215

Paris, D., & Alim, H. S. (2017). *Culturally sustaining pedagogies: Teaching and learning for justice in a changing world.* Teachers College Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* National Academy Press.

Simba Information. (2019). *PK–12 Testing Market Report: PreK–12 instructional materials industry competitive analysis.* Simba Information. https://www.simbainformation.com

Stanovich, K. (1986). "Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy." *Reading Research Quarterly, 21*(4), 360-407.

West, M. R., Morton, B. A., & Herlihy, C. M. (2016). *Achievement Network's Investing in Innovation Expansion: Impacts on Educator Practice and Student Achievement.* Achievement Network. http://cepr.harvard.edu

Assessment as a Catalyst for Identity Development, Skill Cultivation, and Social Impact

Saskia Op den Bosch, Jennifer Charlot, Clarissa Deverel-Rico, and Susan Lyons

Abstract

This chapter explores how RevX's assessment system extends beyond content mastery to nurture identity development, skill-building, and real-world impact. Through the lens of a fourth-grade student, Lana, we illustrate how a structured cycle of action, reflection, and feedback supports learners in developing resilience, critical thinking, and a sense of agency. RevX's DEEDS framework—Discover, Examine, Engineer, Do, Share—guides students through real-world problem-solving, positioning assessment as a tool for growth rather than a static measure of performance.

Rooted in sociocultural learning theories and critical pedagogy, RevX integrates formative and summative assessments to shape responsive instruction, ensuring students see themselves as capable change-makers. By embedding identity-affirming assessments into project-based learning, students not only acquire disciplinary knowledge but also develop the confidence to navigate challenges and contribute meaningfully to their communities. This chapter details how the RevX assessment model—grounded in intellectual prowess, strong sense of self and community, and the ability to create impact—redefines traditional metrics of success. Looking ahead, we discuss ongoing efforts to validate and scale this model, demonstrating how assessment, when intentionally designed, can empower learners to see their own potential and step into their roles as leaders and problem-solvers in an evolving world.

Introduction

A fourth-grade student, Lana, set out to create floor tiles that could harness energy from footsteps, aiming to reduce power consumption for a commonly used school item, like a smartboard or computer. Each week, she tackled the principles of energy flow, circuits, and wiring, demonstrating her understanding through standards-based quizzes called Checks for Understanding. However, as she moved from theory to practice—soldering wires, troubleshooting connections—she faced repeated setbacks, often leaving her frustrated, questioning her abilities, and in tears. At times, she withdrew from her team, needing space to process the challenge independently.

RevX's pedagogical model is predicated on addressing a relevant community challenge. Our approach to *Assessment in the Service of Learning* goes beyond content knowledge, supporting skill-building and identity development by encouraging students to explore who they are and what they're capable of as they engage in real-world challenges. This approach is grounded in three key pillars:

1. Action—the hands-on learning experience itself, which pushed Lana to solve complex, real-world problems;

2. Reflection—weekly academic assessments to confirm understanding combined with personal reflections (through journaling, vlogging, or other forms) on her growth, teamwork, and what additional support from her facilitator (teacher) could help her move forward; and

3. Feedback—from her own data reports, her team, and her facilitator, who provided constructive feedback on technical skills and personal development.

Through this structured cycle of action, reflection, and feedback, Lana was able to confront challenges, build resilience, and develop a clearer sense of her capabilities, not just in terms of knowledge but as a growing individual and teammate.

As the project grew tougher, so did Lana's determination. She began coming to school early and staying late, dedicating extra hours to refining her work and double-checking her thinking. Her presentation, which she had to prepare for the Department of Sustainability in New York City, became a focal point for feedback from her teacher, filled with comments that encouraged her to push through

doubts, deepen her technical knowledge, and reflect on how her reactions to challenges affected her team dynamic.

When the Department of Sustainability and graduate engineering students visited, Lana demonstrated an astonishing grasp of the content—not because she had simply studied it, but because she had learned through failure, course corrections, and real-world application. More importantly, the growth and confidence she displayed left a teacher whom she had the year before remarking, "*I almost didn't recognize her.*"

This process was new for Lana; she wasn't just learning circuits, she was also building confidence in herself as a learner and collaborator, which were becoming essential pieces of her identity. In her weekly reflections, she considered not only her academic progress but also her personal growth. The feedback loop became a mirror, helping her see herself more clearly: her strengths, her areas for growth, and her impact on others. She began to identify as someone who doesn't back down from challenges, recalibrates her timeline for meeting her expectations, and sees setbacks as essential steps in her journey. With each reflection, she focused on two goals: maintaining her confidence and learning requisite technical skills, each effort reinforcing her belief in her own capabilities.

Through this experience, Lana's journey illustrated how *Assessment in the Service of Learning* can support the three critical RevX outcomes: Intellectual Prowess, Strong Sense of Self and Community, and Creates Impact. Although her device ultimately only generated enough energy to power a phone—falling short of her original goal—her learning experience was remarkable. By engaging in a cycle of action, reflection, and feedback, Lana discovered that learning isn't just about achieving perfect results; it's about the process of understanding who she is becoming, building resilience, and finding confidence in her abilities, even when success is partial.

In this chapter, we will explore how RevX's assessment system is embedded within an instructional framework that creates meaningful learning experiences and gathers multiple sources of data to inform educator practice. We will revisit Lana's story throughout the chapter to illustrate each component in action.

RevX Origin Story

RevX, a play on Revolutionary Experiences, was born from urgency and built for transformation. Each of our founders came to see how the education system had conditioned us to doubt ourselves—to shrink in the face of power rather than claim it. We were taught to comply, to play small, and as educators, we found ourselves unintentionally passing those same limitations to the young people in our care.

Then, in 2020, young people started asking, "*What can we do about injustice*? *Will my father die because he is Black*?" They felt powerless. We felt powerless. And we knew we couldn't wait for someone else to solve the problem. We knew education had to be a place where students reclaimed their agency, tested their resilience, and built the skills to shape the world on their terms.

That's why we created DEEDS (Discover, Examine, Engineer, Do, Share)—a framework that makes learning active, relevant, and transformative. When students apply real-world skills to solve pressing challenges, they do not just learn—they develop confidence, critical thinking, and a strong sense of self.

Lana's journey mirrors the very reason RevX exists. Her story is the embodiment of what our founders, Jenn, Alexa, Saskia, realized about education. Traditional learning often teaches Black and Brown students to fear failure, to doubt themselves, to wait for permission instead of claiming their power.

- Jenn was told to leave parts of her identity behind to succeed. She later led national school transformation efforts, built alternative schools for disconnected youth, and designed career-connected learning models that empowered students to bridge academic success with real-world application.
- Alexa was once labeled a "delinquent" for missing school. She now leads New York City's top-performing elementary and middle school, proving that rigorous academics can coexist with student empowerment.
- Saskia was conditioned to believe she "wasn't good at math" after failing a test. She now leads national efforts to redesign assessment systems, ensuring students see learning as a tool for growth, not just measurement.

Like Lana, they faced moments where the system told them they weren't enough. Like Lana, they pushed beyond those limitations. And like Lana, they chose to redefine what success looks like.

Through DEEDS, real-world problem-solving, and a reimagined approach to assessment, RevX is shifting education from compliance to confidence, from passive learning to active change-making. The impact isn't just in the skills students build—it's in the identities they claim, the communities they transform, and the power they recognize within themselves. Because when students like Lana step into their full potential, they do not just succeed—they lead.

Overview of the RevX Learning Model

Today's world faces challenges that demand new ways of thinking, creative problem-solving, and a willingness to act with purpose. At RevX, we believe in preparing young people to tackle these challenges by helping them connect their learning to real-world issues, fostering both academic growth and personal identity development. RevX's approach to learning is deeply rooted in sociocultural theories, emphasizing that knowledge is co-constructed through learners' lived experiences and participation in meaningful, real-world activities (Rogoff, 2003; Vygotsky, 1978). We also draw on critical pedagogy literature to support students' social consciousness and action (e.g., Freire, 2020). Our model is designed to position students as active participants in their learning, fostering skill acquisition, identity development, and social engagement.

RevX Outcomes: Developing Key Competencies

The RevX model prioritizes three key competencies that together cultivate well-rounded learners who are capable of meaningful social impact. The first competency, "Intellectual Prowess," fosters critical thinking, problem-solving, and collaboration. Students demonstrating intellectual prowess ask thoughtful questions that deepen understanding and synthesize diverse sources of information to solve complex problems. This competency aligns with the sociocultural perspective that learning is socially situated. Lave and Wenger (1991) argue that knowledge is constructed through active participation in meaningful social contexts. Similarly, Brown, Collins, and Duguid (1989) emphasize that cognitive apprenticeship, where learners engage in authentic problem-solving activities with peers and facilitators, enhances comprehension and skills development. Through the RevX model, students engage in projects that require them to integrate multiple perspectives and navigate complex challenges.

The second competency, "A Strong Sense of Self and Community," supports students in developing self-awareness, resilience, and empathy. Indicators of this competency include students' ability to recognize their strengths and challenges, as well as their capacity to respect and incorporate diverse perspectives in collaborative tasks. This aligns with research on identity formation in learning. Holland et al. (1998) explain that identity is socially constructed, evolving through interactions with peers, mentors, and communities. Nasir and Hand (2008) further highlight that identity is not static but shaped through engagement in learning environments that require students to take on meaningful roles. RevX intentionally embeds learning within authentic social contexts and collaborative work, ensuring that students not only acquire knowledge but also develop a deeper understanding of themselves and their role within their broader communities.

The third competency, "Creating Impact," prepares students to apply their learning in meaningful ways, positioning them as critical agents of social action and justice. This competency is demonstrated when students work in teams, communicate effectively, gather feedback to improve their work, and develop strategies to solve complex social justice challenges. Freire (2020) argues that education should be a tool for liberation, empowering learners to critically engage with the world and take action against systemic injustices. Similarly, Gutstein (2012) emphasizes the role of critical pedagogy in fostering students' ability to analyze and challenge inequitable structures through their learning experiences. By engaging in collaborative, problem-based learning experiences that center on social impact, students are developing disciplinary expertise alongside their abilities to effect meaningful change in the world.

The Instructional Framework: DEEDS

RevX's DEEDS framework serves as the structural foundation of RevX's instructional model, guiding students through five interconnected phases of learning. The Discover phase emphasizes the identification and exploration of pressing societal challenges within relevant cultural and social contexts. This stage aligns with Vygotsky's (1978) theory that learning is mediated through cultural tools and social interactions, with students constructing understanding through guided exploration.

During the Examine phase, students engage in structured inquiry and collaborative research to analyze root causes and potential solutions. Brown, Collins, and Duguid (1989) advocate for an apprenticeship model in which learners gain expertise

through sustained engagement with mentors and peers. This phase ensures that students critically engage with content rather than passively absorb information.

The Engineer phase involves the design and development of actionable solutions, a process through which students transition from novice to expert roles. This form of participation fosters deep learning as students refine their skills through direct application and iteration (Lave & Wenger, 1991).

In the Do phase, students implement their solutions in real-world contexts. This stage reinforces the notion that learning is an active, situated process, allowing students to engage with authentic audiences and refine their work based on feedback and experience.

Finally, the Share phase prioritizes structured reflection, where students assess their growth and articulate their evolving identities as learners and contributors. Gutiérrez and Rogoff (2003) argue that learning is a culturally mediated process in which individuals construct meaning through dialogue, reflection, and interaction with social tools. Within RevX, this reflective practice enables students to recognize and articulate their own development trajectories.

The RevX Learning Model offers a robust framework for fostering socially situated, identity-driven learning experiences. RevX centers the importance of processes of becoming in addition to its emphasis on knowledge acquisition. Through structured engagement, real-world problem-solving, and reflective assessment, RevX cultivates an educational environment in which students are empowered to shape their own trajectories and contribute meaningfully to their communities. This model thus serves as an exemplar of how sociocultural learning theories can be operationalized to create transformative educational experiences.

The DEEDS framework comes to life through six-to-eight-week instructional modules like Power Up, where students engage in a hands-on, real-world challenge that directly impacts their community. Instead of simply studying energy systems in theory, students step into the role of engineers and problem-solvers, applying their knowledge to design sustainable solutions that address real energy challenges in NYC schools. For Lana, a fourth-grade student in the Power Up module, learning was no longer about memorizing facts—it was about solving a problem that mattered. Like many schools in New York City, hers relied heavily on fossil fuels, consuming large amounts of energy daily. Partnering with the NYC Department of

Education Office of Sustainability and Columbia University School of Engineering, Lana and her classmates were tasked with designing and pitching renewable energy solutions that could reduce energy consumption in their school.

Through Power Up, students moved through the DEEDS framework in a structured, purpose-driven way:

- In Discover, they explored energy transformation and conservation, examining how their own school's energy use contributed to environmental challenges.
- In Examine, they researched renewable energy solutions, analyzed real-world examples, and evaluated how sustainable technologies could be applied in school settings.
- In Engineer, they designed and refined prototypes, such as kinetic tiles that generate electricity from footsteps or bike-powered classroom tools, pushing them to think critically and creatively.
- In Do, they presented their solutions to engineers and sustainability experts, applying their learning in an authentic setting and receiving actionable feedback.
- In Share, they reflected on their experiences, considering both their academic growth and their role in shaping a more sustainable future for their community.

The impact of Power Up extended beyond academic learning—it reinforced the core outcomes of the RevX model. As students engaged in problem-solving and real-world application, they strengthened their Intellectual Prowess, building scientific understanding and technical skills. They developed a Strong Sense of Self and Community, recognizing their ability to contribute meaningfully to their communities. And most importantly, they Created Impact, as their ideas and solutions drove tangible change, making sustainability a priority within their school.

The story of Power Up demonstrates how RevX's DEEDS framework transforms learning into a process of discovery, agency, and action. Through structured engagement, students like Lana do not just learn about the world—they learn to shape it. Scan the following QR code to view this module in action.

RevX Assessment System: A Responsive Design

In viewing assessment as part of a coherent system that includes curriculum and instruction (Black et al., 2011; NRC, 2001; Wilson, 2018)—a system that is grounded in shared, sociocultural views of learning and developmentally appropriate models of disciplinary learning—the RevX approach to assessment has been designed to support the DEEDS framework for curriculum and instruction. The multifaceted RevX assessment system supports the idea that young people learn best when their growth is ongoing, rooted in purpose, and responsive to who they are becoming. In line with contemporary calls for classroom assessments that support more than just academic outcomes (NASEM, 2025), RevX assessments encourage learners to engage deeply, see their progress, and understand themselves. Research supports this approach: formative, real-time assessment improves learning and builds self-confidence (Black & Wiliam, 1998).

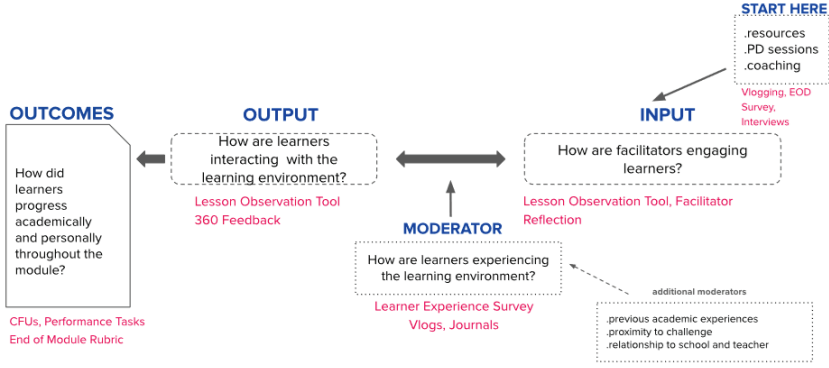
The RevX assessment system seeks to understand and improve each of these focal points individually and the relationship among them:

- Teacher practice: The assessment system aims to ensure that teaching practices build inclusive, engaging, and identity-affirming learning spaces where students feel motivated to learn, valued, and appropriately challenged and engaged (Gay, 2002).
- Learning experience: The assessment system aims to ensure that learning experiences are relevant, challenging, and foster agency, problem-solving, and critical thinking (Bandura, 1986), and to ensure that learners' voices are amplified so that their academic and social-emotional needs are met.
- Learner outcomes: The assessment system aims to ensure that young people are building the foundational knowledge, 21st-century skills and mindsets to step into their roles as community builders.

The RevX Assessment System is built on a theory of action that if facilitators regularly implement quality instruction that is rigorous, relevant, and identity-affirming, and young people engage as intended, then overtime, we will observe academic and personal growth that is increasingly consistent and skilled, across contexts (Figure 1). A facilitator's ability to implement quality instruction will be influenced by the quality of resources, professional development, and coaching provided.

We also hypothesize that the power of facilitator and learner dynamics leading to strong outcomes is effectively moderated by learners' internal states and interpretations of their experiences within the learning environment, which are influenced by their previous academic experiences, their own identity, and proximity to the content, and their relationships to the school, their peers, and their educators.

Figure 1.
RevX Assessment Theory of Action



Integrating Tools for Learning and Growth

At RevX, assessment is not an isolated event but an ongoing process that actively shapes learning, identity development, and real-world skill-building. Rather than treating assessment as a static measure of performance, RevX integrates multiple tools to create a holistic and dynamic feedback system. Tools, like material artifacts and recurring processes, scaffold learning for students and educators (Wertsch, 1988; Stroupe et al., 2019). These tools allow facilitators to monitor student engagement, conceptual understanding, and skill application, ensuring that learning remains responsive and personalized.

As shown in *Figure 1, RevX Assessment Theory of Action*, we hypothesize that when Teacher Practice (e.g., facilitators using high-quality, identity-affirming instructional methods) effectively meets student needs, the Learning Experience becomes more engaging, rigorous, and supportive. In turn, this drives positive Learner Outcomes, such as mastery of disciplinary skills, development of a strong sense of identity,

and the ability to create real-world impact. Conversely, each learner brings prior experiences, motivations, and identities into the classroom, moderating how well the teaching practices land and shaping the learning experience. By gathering feedback on these dynamics, we can continuously refine teaching strategies and support stronger outcomes for every student.

To make this process concrete, RevX uses multiple assessment tools that inform each focal point of the assessment system:

1. ***Learner Experience Surveys*** - Gauge how students feel about their belonging, motivation, and support each day or week. These surveys give facilitators real-time insights into the Learning Experience, revealing how effectively current teaching practices are fostering a supportive environment.
2. ***Checks for Understanding (CFUs)*** - Provide frequent formative snapshots of students' conceptual knowledge and skills. CFUs primarily help teachers fine-tune Teacher Practice by highlighting gaps in understanding. In turn, they also inform Learner Outcomes as teachers adapt lessons to improve mastery and confidence.
3. ***Facilitator Observations*** - Qualitative, real-time notes on classroom interactions, including student collaboration, engagement patterns, and points of confusion. Observational data bridges all three focal points—showing the immediate impact of Teacher Practice on the Learning Experience, and how students' emerging behaviors signal changes in Learner Outcomes or potential areas for intervention.
4. ***Performance Tasks*** - Assess students' ability to apply knowledge and skills in authentic, real-world contexts. These tasks serve as a key indicator of Learner Outcomes, demonstrating students' intellectual prowess and potential for real-world impact. They also feed back into Teacher Practice, helping facilitators refine future instruction.
5. ***Self-Reflections and 360° Feedback*** - Encourage students and peers to articulate growth, challenges, and teamwork dynamics. By capturing student perspectives, reflections and peer feedback illuminate how the Learning Experience is shaping identity development and collaboration. This information loops back to support more responsive teaching and deeper Learner Outcomes.

Processing and Using Data: The Assessment-to-Action Cycle

Each tool corresponds to, and continuously informs, one or more of the focal points in our assessment system: Teacher Practice, Learning Experience, and Learner Outcomes. When used in tandem:

- Teacher Practice can adapt in real time, guided by CFU scores, observational data, and feedback loops.
- Learning Experience improves as facilitators act on data from learner surveys, reflections, and performance tasks, personalizing support for individuals and groups.
- Learner Outcomes become more robust when students receive timely feedback, see relevance in their work, and feel supported in their personal growth and community impact.

This system is built on the belief that learning is iterative. Students must have multiple opportunities to engage with disciplinary content, reflect on their understanding, and receive targeted support. Our facilitators use data reflection protocols and dashboards to analyze patterns, identify student needs, and adjust instruction accordingly. The process begins with key questions: *Are students deeply engaged*? *Are instructional practices supporting identity formation and skill-building*? *Where do students need targeted interventions*?

By systematically applying these assessment tools and analyzing the resulting data, RevX triangulates multiple perspectives—from the learner, the facilitator, and the performance evidence—to paint a comprehensive picture of growth. Rather than viewing assessment as a static, isolated event, we see it as a dynamic, continuous process that both reflects and guides the evolving relationships depicted in Figure 2.

Lana's story illustrates how this approach plays out in practice. Her assessment data revealed early struggles, allowing facilitators to target support interventions that ultimately contributed to her growth.

Lana's Growth: A Story of Progress and Persistence

At the start of the module, Lana's data showed two challenges: Her Check for Understanding (CFU) assessment data revealed she was only scoring 21% on the scientific practice of fair testing (3-5-ETS1-3; NGSS Lead States, 2013) and she

evidenced low engagement, reporting frustration in her Learner Experience Survey. Based on these early CFU scores and classroom observation notes, her facilitator introduced structured experimentation templates and small-group coaching—practical steps to scaffold Lana's troubleshooting process. By pairing her with a peer who excelled at iterative design, her RevX facilitator aimed to increase her exposure to effective problem-solving and build her confidence in a supportive partnership.

Within two weeks, new CFU data (Week 2) indicated Lana still had difficulty connecting *speed to energy transfer,* scoring 25% on 4-PS3-1 (NGSS Lead States, 2013). Recognizing this gap, the RevX facilitator shifted instruction to hands-on ramp-and-ball demonstrations to illustrate how speed affects energy. Brief CFUs with sentence starters prompted Lana to verbalize her thinking, while think-aloud sessions encouraged her to process misconceptions with peers. These focused interventions not only clarified the science concepts but they also seemed to help Lana feel more comfortable voicing questions—a turning point reflected in her Learner Experience Surveys, where she began to report feeling "part of the group."

By Week 3, Lana's ability to generate and compare multiple solutions (3-5-ETS1-2; NGSS Lead States, 2013) improved from 25% at baseline to 50% in that week's CFU, showing she was more open to generating and comparing multiple solutions, though she still struggled to pivot on her designs. Building on that data, the facilitator implemented structured brainstorming sessions with explicit prompts, inviting Lana to explore alternative designs on chart paper. These sessions doubled as a check on her mindset—she could articulate challenges and propose next steps, which in turn gave the facilitator targeted insights on how to guide her.

As summarized in Week 5 data shown in Table 1, Lana's resilience and collaboration were noticeably stronger as observed during a Performance Task, affirmed by peer feedback highlighting her initiative in troubleshooting. Encouraged by these shifts—evidenced by more positive Learner Experience responses—the facilitator provided ongoing one-on-one check-ins and emphasized "small wins" to sustain momentum. Each time Lana demonstrated new problem-solving or collaborative behaviors, the teacher spotlighted them, using immediate feedback to reinforce her growing confidence.

Table 1

Lana's Data Journey Over Time

Time Point	CFU Performance Data	Observation Notes	Learner Experience Data	Data-Driven Instructional Action
Week 1	3-5-ETS1-3 = 21% (Fair Testing & Iteration)	Lana struggled with troubleshooting her prototype; she often withdrew from group discussions after test failures.	Fluctuating sense of belonging; reported frustration with frequent setbacks.	Introduced structured experimentation templates. Paired her with a peer who excelled at iterative design. Held small-group coaching.
Week 2	4-PS3-1 = 25% (Speed & Energy Relationship)	Lana had difficulty connecting the speed of an object to its energy in a preliminary mini-presentation, showing uncertainty about how energy transfers at higher speeds.	Still uncertain about her skills and place on the team; moderate motivation.	Led hands-on "ramp-and-ball" demonstrations to show speed-energy relationships. Used brief CFUs with sentence starters. Encouraged think-alouds.
Week 3	3-5-ETS1-2 = 50% (Generating & Comparing Ideas)	Lana began exploring multiple solutions—though this might reflect greater comfort with brainstorming than with deeper scientific concepts. She still hesitated to pivot her design.	Sense of belonging improved; reported feeling more supported and "part of the group."	Implemented structured brainstorming sessions with explicit prompts. Added reflection journals for analyzing and adjusting her ideas.

Table 1. (continued)

Week 5	Prototype Iterations (Performance Task checks)	Showed stronger resilience and collaboration. Peer and teacher feedback noted she was taking initiative to troubleshoot issues rather than withdrawing.	Reported higher confidence, citing a feeling that "I can figure things out even if it's hard."	Continued 1:1 check-ins and peer feedback loops. Used success milestones (small "wins") to sustain motivation.
Week 7 (final)	Final Pitch & Prototype (Performance Task)	Although Lana's final pitch was overall strong—she demonstrated her working prototype and explained key energy-flow concepts—she still struggled with minor gaps, e.g., detailing how speed affects voltage output.	Reported feeling "very motivated and proud," rating her sense of belonging as consistently high.	Addressed minor clarity issues through last-minute coaching on speed-voltage relationships. Reinforced her progress with positive peer affirmations.

Note. CFU = Checks for Understanding. The final presentation showed that Lana's understanding of energy transfer had improved substantially from Week 1, though she occasionally missed specific cause-and-effect details about speed. Overall, her clarity, confidence, and collaboration were significant leaps from the early stages of the module.

By the final assessment in Week 7, Lana's Performance Task scores indicated she could consistently apply the Science practice of fair testing and explain energy flow (4-PS3-4; NGSS Lead States, 2013). While she still had minor gaps around how speed affects voltage output, targeted last-minute coaching helped refine her final pitch. The NYC Department of Sustainability praised her thoroughness, reflecting both her deeper conceptual mastery and her stronger sense of self. Her Learner Experience data also showed the highest levels of motivation and belonging yet—she reported feeling "very motivated and proud," a testament to how instructional changes, informed by data, had accelerated both her academic and personal growth.

Lana's Growth Was Evident Across All Three RevX Outcomes

- Intellectual Prowess: By Week 7 of the module, Lana had improved her average score on standards-based assessments to around 70%, demonstrating a significant leap in both conceptual understanding and practical application. She progressed from early struggles (Week 1's 21% on fair testing and iteration) to confidently explaining her prototype's energy flow by the final pitch.
- Creates Impact: Although her energy tile prototype did not fully achieve its initial goal—only powering a smartphone rather than a larger device—Lana recognized the value of her learning process. The Department of Sustainability still favored her idea, and engineering students praised the clarity of her explanation about how energy transfer worked through her tile's circuitry.
- Strong Sense of Self and Us: Lana's confidence grew steadily across each week, as reflected in her Learner Experience Survey responses, which indicated rising motivation and sense of belonging. She spoke openly about how her setbacks deepened her self-awareness and collaboration skills. By the time she presented her final work, her self-assuredness was as notable as her improved science comprehension.

By embedding assessment within the learning process, RevX ensures that students like Lana strengthen scientific and engineering concepts and develop the resilience and self-efficacy to thrive in real-world problem-solving. "*The transformation in her self-assuredness was just as remarkable as her improved science understanding,*" her facilitator noted, tying back to the core philosophy underlying RevX's Assessment System design: by using assessment data to shape timely, relevant instructional interventions, educators can help students like Lana reach new heights of competence and confidence—well beyond what simple scores alone would predict.

Educator Training and Supports

Key to the assessment theory of action, RevX ensures that facilitators are equipped with the training, guidance, and resources needed to effectively implement the DEEDS framework and support both academic growth and identity development. Through professional learning workshops, real-time coaching, and data-driven instructional tools, educators learn to interpret assessment data, create identity-affirming spaces, and scaffold student agency.

Facilitators receive structured training on using formative assessments, learner experience surveys, and reflection tools to adapt instruction in real-time. They also engage in ongoing coaching to refine their practice, ensuring every student experiences rigorous, relevant, and empowering learning. By preparing educators to integrate data with student identity development, RevX builds a model that is impactful across diverse learning environments.

RevX recognizes that effective implementation requires more than just training—facilitators need intuitive tools that streamline instruction, assessment, and student support. To enhance consistency and impact, RevX is developing a digital platform that integrates preprogrammed prompts, assessment tools, and module design capabilities, while also capturing and analyzing student data in real-time. This platform will empower educators to implement DEEDS more effectively, ensuring every learner receives high-quality, data-informed, and identity-affirming instruction. Providing digitized on-demand support will also help address sustainability and scalability challenges—discussed in further detail under *Challenges and Strategies for Scaling*.

Connections to the Principles for Assessment in the Service of Learning

RevX's approach aligns closely with the *Principles for Assessment in the Service of Learning*, ensuring that assessments not only measure progress but also support learning, motivation, identity development, and support for individual differences. By integrating formative and summative assessments throughout the learning experience, the RevX assessment system embodies assessment precisely for learning, rather than assessment of learning (e.g., Taylor, 2022; Wiliam, 2011), and provides a structure that nurtures each student's journey of growth, self-awareness, and agency.

Principle 2: Assessment Focus is Explicit and Includes Purposes, Outcomes, Progress Indicators, and Processes that can be Transferred to Other Settings, Situations, and Conditions

RevX assessments are designed not just to measure content knowledge, but to capture progress, competencies, and processes that extend beyond the classroom. The focus on transfer ensures that learning applies to new settings, situations, and real-world challenges.

For example, in the Power Up module, students use scientific inquiry, engineering design, and systems thinking to develop renewable energy solutions for their schools. They analyze energy consumption, prototype alternative power sources, and present their findings to the NYC Department of Education Office of Sustainability and Columbia University engineers.

This aligns with research emphasizing that effective assessments must move beyond isolated academic tasks and engage learners in applying knowledge to authentic, complex contexts. John Dewey (1938) argued that learning should be experiential, connecting knowledge to real-world applications. The ability to analyze, reflect, and act in new situations is a hallmark of deep learning and assessment for transfer.

Through Power Up, students do not just demonstrate an understanding of energy—they develop the confidence and skills to apply their knowledge in different contexts, whether designing sustainable solutions in their communities or advocating for environmental change in the future.

Principle 3: Assessment Design Supports the Learner's Processes, such as Motivation, Attention, Engagement, Effort, and Metacognition

RevX's DEEDS framework ensures that assessments support, rather than hinder, motivation and metacognition. Assessment design must enhance learner engagement, effort, and self-regulation rather than simply measure performance. At RevX, assessments are embedded within learning experiences, allowing students to receive feedback, iterate on their work, and understand their growth trajectory. This aligns with Zimmerman's (2002) research on self-regulated learning, demonstrating that when students can track their progress and set goals, they develop a greater sense of agency and persistence.

Principle 5: Feedback, Adaptation, and Other Relevant Instruction should be Linked to Assessment Experiences

Black and Wiliam's (1998) seminal research on formative assessment highlights the power of continuous feedback in improving learning outcomes, a principle that underpins the RevX approach. The RevX assessment system is designed to provide clear, actionable feedback that informs both students and facilitators of the next steps. Feedback is not just about evaluating past performance—it serves as a catalyst for future learning and decision-making. An integrated dashboard can bring together multiple assessment sources—learner self-reflection, performance tasks, formative assessments, and environmental feedback surveys—to create a holistic picture of student progress. Facilitators use this data to adapt instruction, scaffold learning, and ensure that every student receives personalized support.

RevX's alignment with the *Principles for Assessment in the Service of Learning* demonstrates a commitment to transfer, equity, motivation, and meaningful feedback. Instead of treating assessments as static measures of ability, RevX uses assessments as tools for learning, self-discovery, and social impact. By ensuring that assessments empower rather than restrict learners, RevX is building a model that prepares students not just to succeed academically but to become agents of change in their communities.

Principle 6: Assessment Equity Requires Fairness in Design of Tasks and their Adaptation to Permit their Use with Respondents of Different Backgrounds, Knowledge, and Experiences.

Equity is fundamental to ensuring that assessments fairly measure students' competencies without reinforcing systemic barriers. Assessment equity requires that tasks be culturally relevant, adapted to different backgrounds and experiences, and free from bias. RevX ensures that assessments connect to students' lived experiences and provide multiple ways to demonstrate learning, fostering an inclusive and affirming environment.

- Equity demands differential treatment according to need. RevX achieves this by:
- Designing culturally relevant tasks that resonate with students' diverse experiences,
- Using multiple forms of expression and representation to allow students to demonstrate their knowledge in ways that align with their strengths, and

- Ensuring that assessments are capable of capturing the processes by which abilities are developing.

This commitment to fairness ensures that all learners can meaningfully engage with assessments and that results contribute to better educational opportunities and practices.

Forecasting Future Work for RevX

Ongoing Validation of the Learning and Assessment Model

RevX's next steps focus on validating and refining the DEEDS framework to ensure its effectiveness, adaptability, and scalability across diverse educational settings. Central to this effort is the development of a robust evidence base that is grounded in disciplinary models of learning (Shepard et al., 2018) and a platform that connects student outcomes, instructional protocols, and embedded teacher moves, providing a comprehensive understanding of how the DEEDS framework functions in varied learning environments.

Centering this work in disciplinary models of learning and working from shared definitions of learner experience outcomes supports the ability to establish construct validity. For example, having a deep and detailed understanding of how students might progress within and across grade bands on the performance expectations represented in the Next Generation Science Standards will inform the curriculum and assessment design, along with teacher learning for supporting students in progressing along disciplinary concepts and practices. Similarly, designing assessment and learning experiences that support learner experience will also attend to construct validity if grounded in clear definitions of such outcomes. Gathering evidence—for example—for how the different components of an instructional module attend to and draw on these research-centered definitions would bolster claims for construct validity.

A critical component of this validation is the RevX digital platform—the crux for organizing assessment data and connecting responsive teaching practices—which allows us to easily display data, monitor implementation fidelity, track student progress, and refine instructional approaches in real-time. Investigating how practitioners make sense of and act on these multiple sources of assessment data will provide evidence for validity-in-use or validity related to the consequences of using assessment (Messick, 1998; Shepard, 1997). By capturing and organizing

key learning data, the platform will help educators visualize student growth, engagement, and areas for further development, making assessment a tool for action rather than a static measure of performance. The platform will ensure that the assessment system provides a holistic picture of student learning, triangulating data from multiple sources to offer both a broad and nuanced understanding of progress. By synthesizing performance tasks, formative assessments, learner self-reflections, environmental surveys, and facilitator observations, the system enables educators to see not just what students know, but how they are applying their knowledge, how they experience the learning environment, and how their identity as learners is developing over time. This process ensures that assessment is not fragmented, but instead woven into the fabric of instruction, supporting timely, responsive teaching.

Validation also requires testing the adaptability of the model across different educational contexts. By working with schools in urban, rural, and alternative settings, RevX will study how the DEEDS framework operates in diverse conditions, allowing for refinements that make the model more accessible and scalable—supporting the ability to gather evidence for cultural validity (Solano Flores & Nelson-Barber, 2001). Partnering with educators in these environments will provide valuable insight into how facilitators interpret and implement DEEDS, ensuring that the framework remains flexible enough to meet the needs of varied student and school populations while maintaining fidelity to its core principles.

Another key aspect of validation is the ability to track both immediate and long-term student outcomes. Through future longitudinal studies, RevX will examine how participation in DEEDS-based learning experiences influences not only academic achievement but also identity development, skill transfer, and real-world application. This approach allows for a deeper understanding of how students carry their learning beyond the classroom, reinforcing the idea that education is not just about knowledge acquisition but about shaping capable, confident, and engaged problem-solvers.

Ultimately, this validation process is about more than proving the effectiveness of DEEDS; it is about ensuring that assessment is integrated seamlessly into instruction, making learning more meaningful, identity-affirming, and responsive to student needs. By refining how data is collected, displayed, and used, RevX is working to create an ecosystem where assessment is not just a measure of

past performance but a tool that actively shapes the learning journey, equipping students and educators alike to grow, adapt, and thrive.

Challenges and Strategies for Scaling

As RevX expands, we recognize key barriers to implementation, including resource constraints, varying school contexts, and the need for educator capacity-building. To address these challenges, we are:

- Equipping educators with structured training and coaching to help them integrate DEEDS seamlessly, by providing opportunities to build shared understanding of the theoretical foundations undergirding the RevX approach (perspectives on learning, models of disciplinary learning, definitions of learner experience outcomes), opportunities to make sense of assessment data, opportunities to reflect on appropriate interventions or responsive approaches in light of their students' contexts and needs, even in schools with limited experience in project-based learning;
- Developing an AI-powered digital platform to provide preprogrammed instructional tools, real-time assessment analytics, and adaptive learning supports, reducing the planning burden on teachers and ensuring quality and consistency in implementation; and
- Offering flexible adoption models, allowing schools and organizations to adapt DEEDS in whole and in part to fit their specific needs—whether with just a few strategies or assessment tools, as a standalone program after school, embedded as part of the instructional day, or through a community-based learning initiative.

By proactively addressing these scalability challenges, RevX ensures that its model remains accessible, adaptable, and impactful, creating a clear pathway for schools and communities to implement authentic, student-driven learning experiences at scale.

Long-Term Impact Goals

As RevX grows, we remain committed to empowering young people to actively engage with the world around them and building the capacity of educators, mentors, families, and school leaders to co-design and facilitate these experiences.

By leveraging real-time assessment data, we will support continuous learning, ensuring that both students and educators evolve alongside one another.

Through a sustained focus on student-centered, real-world learning, RevX will continue to refine its model, setting a new standard for education systems that prioritize purpose-driven lives, community engagement, and lifelong growth. Additionally, our research and data collection will contribute to the broader field of education and assessment, offering a replicable model for embedding identity development and real-world learning into assessment practices.

Takeaways for the Field: Assessment as a Catalyst for Identity and Growth

RevX redefines the role of assessment, demonstrating that it can be more than a measure of academic achievement—it can be a catalyst for personal growth, skill development, and social impact. Through the DEEDS framework and its embedded research and development system, RevX integrates assessment into the learning process, making reflection, feedback, and action central to every student's journey. Rather than treating assessment as separate from learning, RevX positions it as a tool to help students recognize their strengths, expand their thinking, and see the impact they can have on the world. This approach has the potential to not only improve academic outcomes but also build agency, confidence, and a deep sense of purpose, proving that assessment—when designed with intention—can be a force for transformation.

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspective*, *9*(2–3), 71–123.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42.

Dewey, J. (1938). *Experience and education*. Macmillan.

Freire, P. (2020). Pedagogy of the oppressed. In *Toward a sociology of education* (pp. 374–386). Routledge.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19–25.

Gutstein, E. (2012). *Reading and writing the world with mathematics: Toward a pedagogy for social justice*. Routledge.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (1998). *Identity and agency in cultural worlds*. Harvard University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Nasir, N. I. S., & Hand, V. (2008). From the court to the classroom: Opportunities for engagement, learning, and identity in basketball and classroom mathematics. *The Journal of the Learning Sciences*, *17*(2), 143–179.

National Academies of Sciences, Engineering, and Medicine. (2025). *Equity in K–12 STEM education: Framing decisions for the future*. National Academies Press. https://doi.org/10.17226/26859

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

National Research Council. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. https://doi.org/10.17226/18290

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–24.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(5), 553–573.

Stroupe, D., Moon, J., & Michaels, S. (2019). Introduction to special issue: Epistemic tools in science education. *Science Education*, 103(4), 948–951.

Taylor, C. S. (2022). *Culturally and socially responsible assessment: Theory, research, and practice*. Teachers College Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard University Press.

Wertsch, J. V. (1988). *The social formation of mind*. Cambridge University Press.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20.

Zimmerman, B. J. (2002). Becoming a self-regulated learner. *Theory Into Practice*, *41*(2), 64–70.

Learning to Read Doesn't End in Third Grade: Supporting Older Readers' Literacy Development with a Validated Foundational Skills Assessment

Rebecca Sutherland, Mary-Celeste Schreuder, and Carrie Townley-Flores

Abstract

Chronically low reading proficiency rates in upper elementary, middle, and high school are a perennial education issue across the United States. Wang et al.'s 2019 investigation of the decoding threshold phenomenon introduced empirical evidence indicating that many older students struggle with reading comprehension because they have inadequate decoding skills. This finding points to a need for current, developmentally appropriate assessment of older students' foundational reading skills, from more advanced skills like morphology knowledge and multisyllabic word recognition, to basic skills like phonics knowledge and phonemic awareness, all of which contribute to older students' reading efficiency, accuracy, fluency, and comprehension of grade-level text. Older students' heterogeneous literacy learning profiles require accurate diagnosis of existing skills and areas of instructional need. The chapter includes a description of the ROAR (Rapid Online Assessment of Reading) a new, free computerized assessment of foundational literacy skills that is validated for use with K–12 students, as well as insights from a pilot initiative that supported middle and high school teachers to administer ROAR to their students and then use the assessment data for instructional planning and progress monitoring.

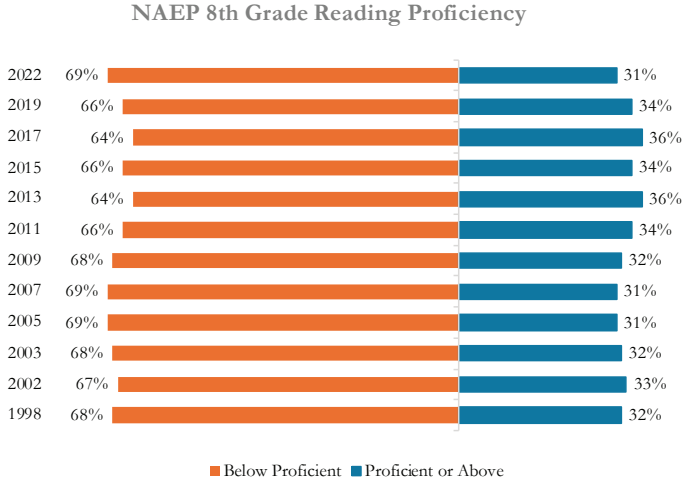
Introduction

Taking a Closer Look at Reading Proficiency Among Older Students

Literacy is the cornerstone of academic success for students in upper elementary, middle, and high school. Across subjects, older students are routinely expected to learn new material through independent reading (Solis, Kulesz, & Williams, 2022; Shapiro, Sutherland, & Kaufman, 2024). And yet, data from the National Assessment of Educational Progress (NAEP) Reading Assessment indicates that this is an unreasonable expectation for the majority of students in upper elementary, middle school, and high school. In 2022, only 33 percent of fourth-grade students and 31 percent of eighth-grade students scored at or above NAEP Proficient level, which is described as "solid academic performance and competency over challenging subject matter" (Nation's Report Card, 2022a & 2022b). These startlingly low reading proficiency rates among older students are also observed in state achievement tests administered annually in districts across the country (Achieve, 2018).

Low reading proficiency rates in upper elementary and secondary school are not a new problem; NAEP assessment data from the last thirty years show consistently flat proficiency rates stretching back to the 1990s (NAEP Reading: National Achievement-Level Results, 2022).

Figure 1.
NAEP Proficiency Chart



The ability to read independently for comprehension is an ultimate goal of reading instruction; students who do not test as "proficient" are likely to struggle to comprehend grade-level texts on their own. While the foundational reading skills of students in kindergarten through third grade are usually measured with dedicated benchmark assessments throughout the school year, in most districts and schools the available data about older students' reading abilities is typically confined to measures of comprehension coming from summative achievement tests administered one time each spring. Year after year, state achievement tests and other standardized tests like the NAEP confirm that sizable majorities of older students cannot read proficiently. Because they primarily measure reading comprehension, these tests offer scarce insight into *why* so many students can't comprehend what they're reading (Tighe & Schatschneider, 2014). In the absence of meaningful assessment data, teachers, parents, and students are left in the dark about what is holding them back from being able to read and comprehend the texts they encounter at school (Valencia & Buly, 2004).

How to account for this collective blindspot? The dearth of up-to-date, accurate measurements of older students' foundational reading skills can be connected to long-held assumptions about how students learn to read (Houck & Ross, 2012). The National Reading Panel's (2000) five pillars of literacy (phonemic awareness, print concepts, phonics/word recognition, fluency, and comprehension) describe the foundational literacy skills that *early* elementary students need in order to both decode and comprehend grade-level texts, reflecting the belief that, "in [grades] K–3 children are learning to read, and in [grades] 4–12 children are reading to learn" (Chall, Jacobs, & Baldwin, 1990). This truism accurately describes Tier I literacy *instruction* in most U.S. schools: explicit instructional support is provided to help the youngest students acquire and apply the early foundational skills that allow them to read and comprehend text up through third grade, and then explicit decoding instruction stops.

But does this instructional norm align with most students' literacy learning needs? At first glance it might seem to since, among older students, research shows that the relationship between reading comprehension and those early decoding skills diminishes; older students' reading comprehension has been found to be more strongly associated with their language comprehension, vocabulary, and background knowledge (Lonigan, Burgess, & Schatschneider, 2018). Why would teachers waste precious class time on unnecessary explicit decoding

instruction? Indeed, Share's (1995) Self-Teaching Hypothesis proposes that, once students have mastered sound-letter correspondences and the essential phonics skills of segmenting and blending, they should be able to independently apply their knowledge to learn novel words. In this view, proficient readers are able to decode and learn unfamiliar words by attending to the order of letters, using their understanding of how the letters map onto oral speech. Until recently the prevailing belief among both researchers and educators has been that students who have mastered basic decoding skills do not need further explicit decoding instruction in order to read and comprehend independently. Accordingly, most upper elementary and secondary schools do not routinely test their students' ability to decode grade-level text.

Crucially, both the Self-Teaching Hypothesis and the broader belief that children learn all the decoding skills they will need in K-3 treat "decoding skills" as a discrete, singular endeavor, with mastery of sound-letter correspondences and basic phonics being what's needed for students to successfully decode texts of increasing length, complexity, and difficulty. This perspective informs which decoding skills are measured in the tests of literacy knowledge that schools use to plan instruction and monitor progress. These tests have also been used by researchers to examine the relationship between decoding ability and independent reading comprehension. Widely used tests such as DIBELS ® (Dynamic Indicators of Basic Early Literacy Skills) assess foundational skills like phonological awareness, rapid automatized naming, alphabetic principle, single-word recognition, and oral reading fluency (University of Oregon, 2018–2019). Such tests provide rich detail about students' early decoding skills. However, the observed disconnect between *early* decoding skills and *older* students' grade-level reading comprehension may well be an artifact of a failure to recognize that there are more advanced decoding skills which older students must bring to bear as they progress to more complex text.

Emerging evidence indicates that early decoding skills, on their own, are necessary but insufficient for older students to achieve and maintain grade-level reading proficiency with texts of increased complexity. A ground-breaking 2019 study utilized an unusual dataset consisting of measurements of upper elementary, middle, and high school students' foundational literacy skills that included not only the standard suite of basic decoding skills that K-3 reading screeners usually test

but also more sophisticated skills that are usually not taught (or assessed) in early elementary grades, like morphology knowledge (Wang, Sabatini, O'Reilly, & Weeks, 2019). These more sophisticated skills are instrumental for decoding more difficult text that students encounter after third grade. Wang et al's analysis revealed a *decoding threshold*, a consistent relationship between older students' expansive decoding skills and their grade-level reading comprehension. Across grades, students whose assessed decoding abilities were below a threshold value tended to have low comprehension scores, while students whose decoding skills were higher than the threshold value on the assessment tended to have stronger (and more variable) comprehension scores (Wang et al., 2019).

Variability in comprehension scores among students who are past the decoding threshold indicates that mastery of basic *and* advanced decoding skills is not a silver bullet that will transform all striving readers into proficient readers who can comprehend grade-level text; some students need support in other critical areas. But, the Decoding Threshold Hypothesis asserts that without adequate decoding skills, older students will not be able to independently read and comprehend grade-level text. This approach was replicated in 2024 with a larger dataset, and the same consistent relationship between students' decoding skills and their ability to comprehend grade-level text was observed (Wang, O'Reilly, & Sutherland, 2024). With growing evidence that, in English, decoding skills continue to undergird reading comprehension beyond third grade, it is time to reconsider how we approach both reading instruction and assessment for older students.

The case for foundational skills assessment in upper elementary, middle, and high school

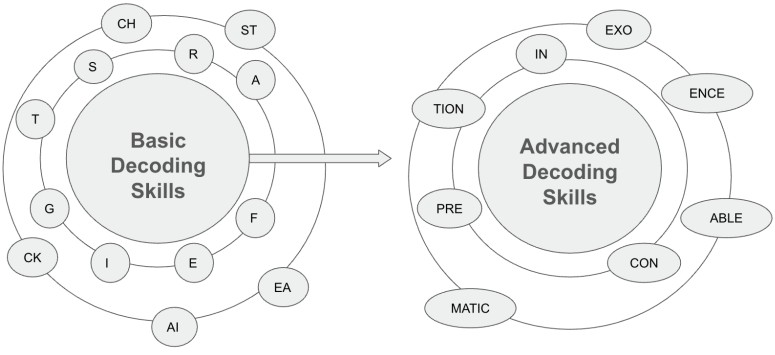
There is increasing heterogeneity in the learning profiles of older readers (Smith & Miller, 2018). To address this variability, a developmentally appropriate foundational skills screening assessment for older students that includes more advanced decoding skills can help teachers to identify and tailor effective instruction that will support individual students to achieve lasting reading proficiency. Accurate, current foundational skills assessment data will allow upper elementary and secondary teachers to differentiate reading instruction appropriately for students with a wide range of literacy support needs, e.g.,:

1. Some students may not have received adequate basic foundational reading instruction in early elementary grades, leaving them ill-prepared to independently read grade-level texts as they move into upper grades. Foundational skill screening assessments will allow educators to quickly identify such students for remedial support.

2. To read and comprehend grade-level texts older students must use more sophisticated decoding skills, including multisyllabic word decoding and knowledge of morphology (Nagy & Anderson, 1984; Tighe & Schatschneider, 2014). Texts in the upper elementary and secondary grades contain novel vocabulary that's often discipline-specific and abstract, along with longer sentences featuring more complicated syntax, and an increasing prevalence of multisyllabic words borrowed from other languages. Words from languages like Greek and Latin have different orthographic rules than what students typically learn in early elementary phonics instruction. Foundational skill screening that tests multisyllabic decoding and morphology knowledge will allow teachers to know which of their students have adequate basic decoding skills but still need explicit instructional support for decoding more complex grade-level text.

3. Students who cannot independently comprehend grade-level text, but have already demonstrated mastery of both basic and more advanced decoding skills, can be appropriately supported in other critical areas, e.g., vocabulary and background knowledge.

Figure 2.
Basic & Advanced Decoding Skills Illustration



When equipped with the right data that pinpoints where individual learning needs lie, upper elementary, middle, and high school educators can support their students to proficiently read and comprehend grade level text.

The remainder of this chapter will describe the development and features of the Rapid Online Assessment of Reading (ROAR), an online screening assessment of foundational literacy skills designed for students K–12, and early lessons drawn from The Achievement Network's (ANet) pilot initiative to implement the ROAR assessment to middle and high school students in Franklin County Schools (pseudonym)—a small urban district in the Northeast United States that has long struggled with low reading proficiency among its older students.

A Validated Foundational Skills Assessment for Older Readers: The Rapid Online Assessment of Reading (ROAR)

The Rapid Online Assessment of Reading (ROAR) emerged from more than a decade of research in Stanford's Brain Development & Education Lab on the neurobiological foundations of literacy overall, and in particular on the brain-based etiology of different subtypes of dyslexia. Identifying difficulties consistent with dyslexia requires measuring key foundational reading skills, which is why these skills are included in the ROAR Foundational Reading Skills Suite. It quickly became apparent that the initial set of ROAR assessments could have utility beyond the world of lab-based research, as they provide accurate, relevant measures of literacy skills that educators can directly use to plan instruction. With all subsequent research and development on ROAR being done in partnership with schools across the country, ROAR bridges the school, community, and lab. Leveraging the extensive literature on the cognitive neuroscience of reading development, the team responded to the needs voiced by school partners by developing an automated, lightly gamified online assessment platform that could replace the resource-intensive and time-consuming conventional approach of individually administering assessments that are scored based on verbal responses. The ROAR platform can assess an entire school system in the time typically required to administer an assessment to a single student. In ten minutes, a teacher can assess a classroom on word-level decoding and sentence reading efficiency to evaluate the risk for reading difficulties such as dyslexia. In 45 minutes, ROAR can provide a more detailed profile of strengths and areas needing support. ROAR can be administered to all students just once to screen for skill mastery, or multiple times throughout the

year for progress monitoring of targeted skill areas. ROAR can be used across the grades, K–12, filling a gap in older grades where screening and progress monitoring for foundational reading skills is strongly needed but historically neglected due to a lack of time, resources, teacher training, and available assessments.

ROAR consists of a collection of measures, each designed to assess a specific domain of reading. Each measure can be run independently and returns scores to teachers in real time. ROAR is designed as a series of assessment modules that can be sequentially administered in one sitting or individually. ROAR assessment modules test students' foundational literacy skill knowledge, including:

- Phonemic Awareness
- Letter Naming
- Letter Sound Knowledge
- Phonics Knowledge (2026 release)
- Single Word Reading
- Sentence Reading Efficiency
- Morphology
- Syntax
- Inference
- Vocabulary

Core assessments are also available in Spanish. Across the country, the ROAR team is collaborating with schools to understand how foundational reading skills assessments in multiple languages may combine to support targeted intervention for multilingual learners including newcomers and long-term English learners.

Pushing the frontier of reading assessment, the ROAR team is working alongside schools to research how the integration of rapid automatized naming, visual processing, and executive functioning measures alongside measures of foundational reading skills may support more targeted interventions for dyslexia and other reading issues.

Dedicated to the design principle of assessment quality, which includes utility, credibility, and making appropriate inferences, ROAR measures are designed to be user-friendly for both teachers and students. ROAR measures are also individually assessed for reliability, concurrent validity, and predictive validity.

The ROAR Technical Manual provides these statistics by grade, race, ethnicity, gender, special education status, English learner status, and free lunch eligibility. ROAR measures are strongly correlated ($r > 0.8$) with gold-standard measures such as the Woodcock-Johnson, Comprehensive Test of Phonological Processing (CTOPP), Test of Word-Reading Efficiency (TOWRE), and Test of Silent Reading Efficiency and Comprehension (TOSREC) (Yeatman, Townley-Flores, et al., 2024; Yeatman, Tran, et al., 2024; Yeatman, Tang, et al., 2021; Gijbels, et al., 2024). These robust psychometric properties ensure that ROAR provides educators with reliable and equitable data to support informed decision-making and effective instruction across diverse student populations.

What should a district do to prepare for success when adopting a foundational reading skills assessment tool for older students?

Adopting a new assessment tool for older grades presents significant challenges. Teachers are already burdened with extensive classroom demands, a situation exacerbated by the pandemic (Jomuad et al., 2021). On average, older students spend 20–25 hours per school year taking state- and district-mandated assessments (Jimenez & Boser, 2021). This underscores the importance of adhering to the *principles of assessment in the service of learning*. Effective assessments should provide transparency for all stakeholders, offer actionable feedback to guide decision-making, and include clear next steps. Additionally, the design of an assessment must support the learner and demonstrate high quality and validity.

However, many districts seeking to assess foundational skills in older students face two key issues: they either use assessments that are not validated for middle or high school students or rely on tests that fail to measure the specific skills required for proficient reading, such as using comprehension assessments to screen for foundational skills. These missteps contribute to assessment and data overload for teachers, particularly when attempting to integrate new tools like ROAR into an already-packed schedule. ROAR addresses these challenges by offering a rapid and automated assessment that can evaluate an entire class in as little as ten minutes, minimizing disruption and maximizing efficiency.

Through ANet's experience piloting ROAR in middle and high schools, we have identified three critical challenges to addressing foundational reading skills in

secondary schools. We believe every school/system leader should consider these challenges when adopting a foundational reading skills plan for older students:

- Creating teacher buy-in for a new assessment and intervention system
- Aligning on a multitiered goal-setting and communication plan across leadership in systems and schools
- Providing districts and schools support in analyzing and taking action on their data through professional development and selection of intervention curricula for students with the highest needs.

Challenge: Achieving Teacher Buy-in by Addressing Common Beliefs

One hurdle often encountered when adopting a new assessment tool is the beliefs of school leaders and teachers. The mindsets of the faculty and staff play a vital role in successful implementation (Laine & Tirri, 2023). When confronted with a new school initiative, there is frequent resistance to change stemming from comfort with current assessments, fear of the unknown, and concerns over the work involved (Lomba-Portela, 2022). While such resistance is understandable, developing a clear purpose for the assessment and interrogating teacher beliefs is crucial.

One belief that may prove to be a hurdle is the notion that early education and elementary teachers alone bear responsibility for supporting foundational skills. While reading must be taught in the younger grades, older students will always need this support as well. For the adoption of ROAR to take hold, teachers of older students must accept their own responsibility for their students' foundational reading skills.

Teachers may also believe that they will have to sacrifice to make room for new practices. Again, this is a valid concern. With any new initiative comes work and the requirement of making space in an already packed curriculum. That being said, if a strong enough purpose is built, teacher responsibility for the success of their students will take precedence over the challenge of making room for new types of instruction. Based on research, foundational skills strategies must be used in the tier 1 classroom, as well as in tier 2 and 3, with complex, grade-level texts (Swanson et al., 2017). Older students must not miss out on their general education classes in favor of interventions. Instead, they need both.

This then leads to a final hurdle—the mistaken belief that making a shift toward foundational skills strategies will be detrimental for students reading on grade level. In Anet's ROAR pilot work, we heard criticism from leaders about the consequences for proficient readers if foundational skills practices are implemented in the tier 1 classroom. This belief stems from the idea that proficient readers have nothing to learn from foundational skills practice and will stagnate if not intellectually pushed. In reality, foundational skills strategies are for all students, not just those who experience challenges with reading. For example, activities such as morphological word work and oral reading fluency practice not only support striving readers but also enhance the reading skills of those who are already proficient (James et al., 2021; White et al., 2021). Adopting an assessment tool, like ROAR, enables leaders and teachers to track this type of growth in all students.

Solution

Prior to adopting a new protocol for addressing foundational reading skills in older students, it is crucial to set aside time in professional development to build up teachers' knowledge of the assessment and develop their mindsets around their role in addressing these skills.

Developing buy-in must begin when stakeholders learn about the assessment's adoption. This involves clearly articulating the purpose and goals for the initiative, presenting the research behind the assessment's efficacy, and sharing success stories from other schools utilizing the assessment. In particular, testimonials are a powerful way to humanize the initiative and demonstrate its relevance to daily work and professional growth. When teachers understand the positive impact of the work, they will be more motivated to put forth the necessary effort for a new assessment.

After buy-in is established, teachers also need training to learn to administer the assessment and analyze the data. If the school does not have a recurrent and designated time for teacher professional learning, it may prove difficult to provide the information necessary to successfully adopt a new foundational reading skills assessment.

Vignette

In the early phases of ROAR's development in Franklin, we struggled to recruit ELA leaders and teachers in the pilot. This was in part due to challenges with communication, but it also stemmed from an underdeveloped purpose. Teachers believed their older students were struggling with reading, but did not see themselves as part of the solution. Instead, they expressed that change first needed to happen at the district level before anything could alter in classrooms. While the district aspired to highlight ROAR's potentially positive impact on student reading outcomes, it was too little too late. Further eroding teacher buy-in, we found that many teachers struggled to administer ROAR due to a lack of effective training; this then led to longer proctoring times and frustration. Training for ROAR may have felt like an unnecessary burden for teachers upfront, but in the long term, it would have alleviated unnecessary snags in the adoption process.

Learning from this, in our second year of the ROAR pilot, we planned a series of professional learning sessions. When starting any new initiative in an educational context, ongoing professional development and support are essential. In fact, professional development is one of the most powerful tools districts have to enhance teacher effectiveness (Hirsh, 2017). For a strong implementation of an assessment, professional learning must happen regularly and be structured around the latest research and most relevant content (Savitz et al., 2024). In the ROAR pilot, we offer up to five professional development sessions focused on ROAR data. Ideally, these sessions are conducted in person with school leaders and teachers, but they can also be offered virtually. The sessions follow a specific schedule tied to the administration of the ROAR assessment. The first professional learning session takes place at the beginning of the school year before the initial ROAR assessment administration. It provides information about older students and foundational skills instruction in middle and high school, as well as a kick-off to the ROAR assessment where we establish a strong purpose for the initiative. The subsequent PL sessions occur 2–4 weeks after each ROAR administration, allowing leaders and teachers time to review the data and formulate questions before engaging in the PL. During these PL sessions, we analyze the data sets and determine the necessary next steps for instruction and intervention to support students. Specific strategies are taught that teachers can immediately implement, and they then bring their classroom experiences and data back to the next PL. As a result, professional learning becomes a collaborative community where participants share their challenges, successes, and artifacts from the implementation cycles of foundational skills strategies for their students.

Challenge: Objectives and Communication Alignment Between System-Level and School-Level Leaders

A strong rollout of a new assessment can substantially influence the acceptance and sustained utilization of such assessment. This involves alignment between district and school leaders on the overall objectives and goals for the adoption and use of the assessment. From our experience with rolling out ROAR in ANet's pilot programs, some district leaders struggle to understand the purpose of different types of literacy assessments, and they use these assessments interchangeably, resulting in inappropriate data application. According to the principles for assessment, assessments should be transparent, with a clear evaluation process and purpose. As an example, a comprehension assessment, such as NWEA MAP Growth, should not be used to determine which students need foundational skills support. In much the same way, ROAR should only be used to screen students for potential gaps in their foundational literacy skills, not to diagnose the discrete skills needing extra support. Once leaders understand the purpose of the assessment, they can then set specific, measurable objectives and goals to guide the implementation process. To align and develop these strong goals, leaders should ask themselves:

- Who is the intended audience for the assessment?
- How will we use the assessment data to drive instructional decisions and support students? What do we expect others, such as teachers, to do with the data?
- When do we expect to see measurable student growth on the assessment, and what targeted instructional strategies will we implement to get there?

Collaborating on the answers to these questions moves leaders one step closer to a smooth implementation of the new assessment. However, goals are not enough to create alignment between the district and school leaders and teachers; communication between a variety of stakeholders also requires attention.

Solution

In the first year of assessment adoption, it is important to establish a working group comprising district leaders, ELA instructional specialists, and teachers who are tasked with developing a strong communication plan for the assessment implementation. This involves strategically determining the sender and audience for each type of communication, selecting the most effective methods for communication, and establishing a timeline. Importantly, the core message of each communication must be clear and specific, providing the right information at the right time. By involving multiple stakeholders in the communication process through the working group, the messaging around the new assessment is not top-down; rather, it is a collaborative effort among colleagues, fostering a shared responsibility for the successful adoption of the assessment.

Vignette

In our first year working in Franklin, we failed to develop a strong communication plan, resulting in haphazard messaging about the purpose of the ROAR assessment. Consequently, school leaders were skeptical about ROAR and saw it as just one more item on their already overburdened "to-do" list. In Franklin's second year, we learned from the challenges of Year 1 and created a working group as described above. Thus far, this has led to a smoother rollout and an enthusiastic reception by school leaders and teachers who understand the purpose and promise of ROAR and subsequent interventions in their middle and high schools.

Challenge: Analyzing the Data and Acting on It

Data must never be for the sake of data collection. As is mentioned in the principles for assessment, the feedback from an assessment must lead to actionable insights for teachers and educational stakeholders that result in the betterment of student learning. For this to take place, educators need support to engage with novel data. One common challenge for secondary educators is determining feasible instructional moves they can take to support their students based on assessed areas of need. The root of this issue harkens back to the research on secondary ELA teachers needing to be trained in reading instruction and receiving minimal professional development in supporting their older striving readers (Moats, 2020). Without the knowledge of evidence-based instructional moves to enhance reading, teachers are left to fend for themselves, armed with comprehension strategies that will not move the needle for students who are scoring below the decoding threshold (Wang et al., 2019). Teachers also need time and support to turn the

data into actionable insights that help them make instructional decisions. These instructional decisions are usually differentiated into tiers of support, with tier 1 support happening at the classroom level, tier 2 in small groups, and tier 3 the most targeted, intensive, and often one-on-one support.

Solution

Students categorized as "Need Some Support" for foundational reading skills on the ROAR assessment require a blend of literacy instruction to develop their decoding and/or fluency skills; this involves tier 2 small group support. These students can be grouped based on their specific needs and provided with differentiated instruction during tier 1 class time (Rasinski, 2017). For instance, while some students work collaboratively to analyze the meaning of complex, multisyllabic words in their text, the teacher can 'push in' to support a smaller group of four to six students whose scores indicate a need for focused decoding instruction. During this push-in support, the teacher could work through the phonemes, syllabication, or morphology of the same words the other students are addressing in their peer groups. The selected students would receive more targeted teacher attention and the opportunity to practice and ask questions in a small group setting. The advantage of push-in support in middle and high school is that older students have greater autonomy and can work in their own groups with minimal supervision, freeing the teacher to support a select group (Jones, Conradi & Amendum, 2016).

Students categorized as "Need Extra Support" on ROAR should be placed in the right tier 3 intervention based on their area of need: decoding or fluency. However, this is not always easy in the secondary setting. As opposed to elementary, middle schools and especially high schools often lack the flexibility in their schedules for an intervention block. This is often due to the amount of credits students need to graduate, which doesn't take into account the potential need for foundational reading interventions. To address this issue, some schools have introduced a 'reading remediation' class that takes the place of students' tier 1 ELA class. However, this is not an acceptable solution. When older students are removed from tier 1 ELA instruction, they miss out on vital content learning as well as experience with grade-level complex texts. Older students need a blend of literacy learning while their decoding and/or fluency needs are addressed (Vaughn & Fletcher, 2021). We recommend system-level coaching to support district leaders in redesigning students' instructional time.

We also recommend that system-level leaders conduct an audit of the literacy intervention programs currently in use in their secondary schools. This process, coupled with insights from ROAR data, may reveal the need for higher-quality materials to support tier 3 interventions. Unfortunately, many available programs for older students are ill-suited, relying on overly simplified gamification and content that does not align with the maturity of teenage learners. To address this, districts must allocate resources to adopt instructional tools and materials that enhance decoding and fluency, which are essential components for meeting the needs of striving readers.

In many districts ANet partners with, multiple intervention curricula are implemented with little evidence from assessment data of their effectiveness. When these programs fail to meet students' needs, leaders must identify the most effective intervention curriculum for improving decoding and fluency in older students and collaborate with teachers to ensure its consistent implementation. This highlights the importance of not only selecting the right curriculum but also equipping teachers with the tools and support they need to adeptly analyze assessment data to make the best decisions about implementing intervention strategies and curricula.

Vignette

In the case of the ROAR assessment, data is relatively easy to understand once technical knowledge is built. In Franklin, we offered targeted training sessions to equip educators with the skills to utilize and analyze the ROAR data. These sessions were one hour in a virtual setting and facilitated by the ROAR lead and coaches from ANet. Educators were given time and support in accessing their school's data along with hands-on instructions for filtering data in numerous ways to offer more specific insights. For the analysis of data, ANet provided a protocol for moving through the data systematically in order to develop best practices for data interpretation. These virtual sessions allowed for collaboration between educators from different schools in order to share insights and discuss common challenges. Educators then dispersed into smaller breakout rooms to work one-on-one with their coach to organize their individual school's data and practice filtering, analyzing, and gleaning actionable insights. Even after these virtual sessions, coaches continued to work with their school leaders and ELA educators to practice data-driven decision-making for instructional change.

Teachers must be aware that data analysis is simply the beginning of any new assessment implementation; it cannot solve the problem of unmet learning needs. Once data is collected and analyzed, action is needed to create any real and lasting change for student learning.

Conclusion

Supporting teachers to support older students' literacy development

Understanding the larger continuum of decoding skills that are required to read and comprehend texts of increasing length and difficulty, paired with assessment data that accurately measures older readers' foundational literacy skills, will reveal where students in upper elementary, middle, and high school need explicit reading instruction. However, the assessment data itself will not provide the instructional support that older students need. Foundational skill instruction that meets students' individual learning needs is only possible when teachers are trained and resourced to both engage with accurate, developmentally appropriate literacy assessment data and to use that data to identify and deploy appropriate instruction (Basma & Savage, 2023).

A majority of upper elementary and middle-school teachers currently report that they have not received relevant pedagogical training to support their students' literacy development; moreover, a large majority of teachers reported that they do not have adequate access to developmentally appropriate instructional resources to support older students (Shapiro, Sutherland, & Kaufman, 2024). Meeting older readers' unrecognized foundational literacy learning needs will require a paradigm shift in how we approach reading instruction—one that acknowledges the broader range of foundational skills students need to read and comprehend increasingly complex grade-level texts, while also providing teachers with developmentally appropriate training and resources. Decades of flagging reading proficiency rates point to the urgency of making this shift.

References

Achieve. (2018). *Proficient vs. prepared: Disparities between state tests and the 2017 National Assessment of Educational Progress (NAEP).* https://www.achieve.org/files/Proficient%20vs.%20Prepared%20May2018_1.pdf

Applegate, A., Applegate, M., McGeehan, C., Gibbons, C., Norris, M., Doyle, K., & Romani, A. (2022). What has changed in state reading tests in 10 years? The NAEP study revisited. *Yearbook of the College Reading Association*, *43*, 113–126.

Basma, B., & Savage, R. (2023). Teacher professional development and student reading in middle and high school: A systematic review and meta-analysis. *Journal of Teacher Education*, *74*(3), 263–278.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Harvard University Press.

Gijbels, L., Burkhardt, A., Ma, W. A., & Yeatman, J. D. (2024). Rapid online assessment of reading and phonological awareness (ROAR-PA). *Scientific Reports*, *14*(1), 1–16. https://doi.org/10.1038/s41598-024-51784-0

Houck, B. D., & Ross, K. (2012). Dismantling the myth of learning to read and reading to learn. *ASCD Express*, *7*(11). https://www.ascd.org/el/articles/dismantling-the-myth-of-learning-to-read-and-reading-to-learn

Hirsh, S. (2017). *Building professional development to support new student assessment systems.* Learning Forward. https://learningforward.org/wp-content/uploads/2017/08/building-professional-development.pdf

Institute of Education Sciences. (2022). *What Works Clearinghouse: Providing reading interventions for students in grades 4–9* (NCEE 2022–005). U.S. Department of Education, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/WWC-practice-guide-reading-intervention-full-text.pdf

James, E., Currie, N. K., Tong, S. X., & Cain, K. (2021). The relations between morphological awareness and reading comprehension in beginner readers to young adolescents. *Journal of Research in Reading, 44*(1), 110–130. https://doi.org/10.1111/1467-9817.12316

Jimenez, L., & Boser, U. (2021). Future of testing in education: The way forward for state standardized tests. *Center for American Progress*. https://files.eric.ed.gov/fulltext/ED617055.pdf

Jomuad, P. D., Antiquina, L. M., Cericos, E. U., Bacus, J. A., Vallejo, J. H., Dionio, B. B., Bazar, J. S., Cocolan, J. V., & Clarin, A. S. (2021). Teachers' workload in relation to burnout and work performance. *International Journal of Educational Policy Research and Review, 8*(2), 48–53. https://doi.org/10.15739/IJEPRR.21.007

Jones, J. S., Conradi, K., & Amendum, S. J. (2016). Matching interventions to reading needs: A case for differentiation. *The Reading Teacher, 70*(3), 307–316.

Laine, S., & Tirri, K. (2023). Literature review on teachers' mindsets, growth-oriented practices and why they matter. *Frontiers in Education*, *8*, Article 1275126. https://doi.org/10.3389/feduc.2023.1275126

Lomba-Portela, L., Domínguez-Lloria, S., & Pino-Juste, M. R. (2022). Resistances to educational change: Teachers' perceptions. *Education Sciences, 12*(5). https://doi.org/10.3390/educsci12050359

Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the simple view of reading with elementary school children: Still simple after all these years. *Remedial and Special Education, 39*(5), 260–273. https://doi.org/10.1177/0741932518764833

Moats, L. C. (2020). *Reading is rocket science.* American Federation of Teachers. https://www.aft.org/sites/default/files/moats.pdf

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304–330. https://doi.org/10.2307/747823

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development.

National Center for Education Statistics. (2022a). *2022 reading state snapshot report: National and State—Grade 4 public schools*. Institute of Education Sciences.

National Center for Education Statistics. (2022b). *2022 reading state snapshot report: National and State—Grade 8 public schools*. Institute of Education Sciences.

National Center for Education Statistics. (2022). *NAEP reading: National achievement-level results*. https://www.nationsreportcard.gov/reading/nation/achievement/?grade=8#nation-achievement-overall

Opatz, M. O., & Kocherhans, S. (2024). Using a supplemental, multicomponent reading intervention to increase adolescent readers' achievement. *Journal of Adolescent & Adult Literacy, 67*, 294–302. https://doi.org/10.1002/jaal.1333

Rasinski, T. V. (2017). Readers who struggle: Why many struggle and a modest proposal for improving their reading. *The Reading Teacher*, *70*(5), 519–524. https://doi.org/10.1002/trtr.1533

Savitz, R. S., Morrison, J. D., Brown, C., Aldrich, C., Kane, B. D., & O'Byrne, W. I. (2024). Secondary teachers' adolescent literacy efficacy and professional learning considerations. *Reading Research Quarterly*, *59*(1), 107–123. https://doi.org/10.1002/rrq.521

Shapiro, A., Sutherland, R., & Kaufman, J. H. (2024). *What's missing from teachers' toolkits to support student reading in grades 3–8: Findings from the RAND American Teacher Panel*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA3358-1.html

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*, 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Smith, H. D., & Miller, S. D. (2018). Digging deeper: Understanding the reading and motivational profiles of students who do not demonstrate proficiency on state-mandated reading assessments. *Reading Psychology*, *39*(8), 855–878.

Solis, M., Kulesz, P., & Williams, K. (2022). Response to intervention for high school students: Examining baseline word reading skills and reading comprehension outcomes. *Annals of Dyslexia*, *72*, 324–340. https://doi.org/10.1007/s11881-022-00253-5

Swanson, E., Stevens, E. A., Capin, P., Scammacca, N., Stewart, A., & Austin, C. (2017). The impact of Tier 1 reading instruction on reading outcomes for students in grades 4–12: A meta-analysis. *Reading and Writing*, *30*(8), 1639–1665. https://doi.org/10.1007/s11145-017-9743-3

Tighe, E., & Schatschneider, C. (2014). A dominance analysis approach to determining predictor importance in third, seventh, and tenth grade reading comprehension skills. *Reading and Writing*, *27*(1), 101–127. https://doi.org/10.1007/s11145-013-9435-6

University of Oregon. (2018). *Dynamic indicators of basic early literacy skills* (8th ed.). https://dibels.uoregon.edu

Valencia, S. W., & Buly, M. R. (2004). Behind test scores: What struggling readers really need. *The Reading Teacher*, *57*(6), 520–533.

Vaughn, S., & Fletcher, J. M. (2021). Identifying and teaching students with significant reading problems. *American Educator*. https://files.eric.ed.gov/fulltext/EJ1281906.pdf

Wang, Z., O'Reilly, T., & Sutherland, R. (2024). *Replicating the decoding threshold in ReadBasix®: Impact on reading skills development* (Research Memorandum No. RM-24–06). ETS.

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology, 111*(3), 387–401. https://doi.org/10.1037/edu0000302

White, S., Sabatini, J., Park, B. J., Chen, J., Bernstein, J., & Li, M. (2021). *The 2018 NAEP oral reading fluency study*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., Caffarra, S., Raikin, S. M., & Richie-Halford, A. L. (2021). Rapid online assessment of reading ability. *Scientific Reports*, *11*(1), Article 6396.

Yeatman, J. D., Townley-Flores, C., Wentzlof, K., Ma, W. A., Siebert, J. M., Fuentes-Jimenez, M., Saavedra, A., Murray, T. S., Bhat, K., Ramamurthy, M., & The ROAR Developer Consortium. (2024). *Rapid online assessment of reading (ROAR) technical manual*.

Yeatman, J. D., Tran, J. E., Burkhardt, A. K., Ma, W. A., Mitchell, J. L., Yablonski, M., Gijbels, L., Townley-Flores, C., & Richie-Halford, A. (2024). Development and validation of a rapid and precise online sentence reading efficiency assessment. *Frontiers in Education*, *9*, Article 1494431. https://doi.org/10.3389/feduc.2024.1494431

A Skills-Based Vision for Assessment, Insight, and Educational Improvement

Ou Lydia Liu, Lei Liu, David Sherer, and Paul G. LeMahieu

Abstract

Our current educational system prioritizes traditional academic disciplines and views the K–12 classroom as the major learning environment. By focusing solely on academic learning, the system overlooks the broader variety of skills learners acquire both inside and outside the classroom, leaving critical skills such as communication, collaboration, and critical thinking underdeveloped. Furthermore, the current approach fails to reflect the diverse pathways through which learners develop expertise, such as military service, internships, or community engagement. Skills-or competency-based education shifts the emphasis from certifying classroom-instilled academic knowledge to certifying students' knowledge and skills gained from a variety of educational, occupational, and societal experiences. This chapter articulates design principles for educational assessments that address a skills focus and meet both academic and workforce needs. Beginning with a review of existing skills frameworks that outline key skills, competencies, and learning outcomes across various contexts in K–12, postsecondary, and workforce sectors, we identify skills deemed critical for the future by looking for commonalities across skills frameworks and state Portraits of a Graduate (PoG) frameworks. After establishing a taxonomy, the chapter discusses how to leverage technology and AI tools to capture skills acquisition, particularly skills that are developed and demonstrated in non-academic context. Then, the chapter discusses assessment design principles that enable the measurement of complex skills with validity, reliability, and authenticity. Finally, the chapter proposes a professional development model and continuous improvement approach that supports the implementation of skills assessment in classrooms.

Keywords: Skills-based assessment, Carnegie unit, multi-modal assessment, continuous improvement, durable skills

Author Note
We have no known conflict of interest to disclose.
Correspondence concerning this article should be addressed to Ou Lydia Liu, Educational Testing Service, 660 Rosedale Rd, Princeton, NJ 08540, United States. Email: lliu@ets.org

A Skills-Based Vision for Assessment and Educational Improvement

The way in which the U.S. educational system credits and validates learning is outdated. Our current system prioritizes traditional academic disciplines and views the K–12 classroom as the major learning environment (Silva et al., 2015). However, this approach fails to reflect the diverse, nonlinear pathways through which learners develop expertise, such as military service, internships, apprenticeships, volunteerism, and community engagement (Werquin, 2023). Furthermore, by focusing solely on academic learning, the system overlooks the broader variety of skills learners acquire both inside and outside the classroom, leaving critical skills such as communication, collaboration, and critical thinking underdeveloped (National School Boards Association, 2025). Most degrees are awarded based on acquisition of academic knowledge, but this narrow focus has led to serious skills gaps among learners (García-Chitiva, 2024; Ulloa-Cazarez, 2021). For example, while close to 100% of employers believe that critical thinking, problem solving, and teamwork are essential skills for workforce performance, less than 60% think college graduates are equipped with these skills (National Association of Colleges and Employers, 2019). To meet the needs of the modern workforce and society, shifting the focus of the U.S. education system to nurture the "whole student" (e.g., Darling-Hammond & Cook-Harvey, 2018; Durlak et al., 2011) is critical to securing the long-term civic and economic flourishing of the country.

In addition to broadening the skills that should be considered in talent preparation, it is also important to expand the pathways through which these skills are acquired. *Skills- or competency-based education*[1] shifts the emphasis from certifying classroom-instilled academic knowledge to certifying students' knowledge and

1 Skills and competencies are often used interchangeably in educational and occupational settings, and we do so as well throughout this chapter.

skills gained from a variety of educational, occupational, and societal experiences. Such a system is agnostic to where students acquired their skills. The focus is on the outcome—demonstrated ability—not the process through which it is developed. Recognizing learning gained through nontraditional pathways allows individuals a wider range of opportunities to demonstrate their qualifications, achieve upward economic mobility, and contribute to society (Bell, 2016).

Existing efforts in competency-based education (CBE) reflect a significant shift toward mastery of skills, competencies, and knowledge through applications in real-world situations. Such a shift is becoming increasingly prominent across K–12 (e.g., XQ Institute, Aurora Institute; Levine & Patrick, 2019), postsecondary (e.g., Western Governors University, 2019; Southern New Hampshire University, n.d.), and workforce sections (e.g., Opportunity@Work). In the K–12 space, organizations such as the XQ Institute and the Aurora Institute have been at the forefront of promoting CBE models, emphasizing personalized learning pathways, allowing students to progress at their own pace once they demonstrate mastery of a given skill of competency. Schools that adopt CBE models are exploring the replacement of traditional grading systems with skill-based assessments. In postsecondary education, institutions like Western Governors University and Southern New Hampshire University have embraced CBE to support adult learners by offering programs where students earn degrees by demonstrating mastery of competencies, rather than accumulating credit hours. These innovative programs allow students to leverage prior experiences from both academic and nonacademic settings to accelerate their skills development. Finally, in the workforce sector, initiatives like Opportunity@Work are reshaping how talent is recognized by advocating a "skills-first" hiring approach, where employers value demonstrated competencies over traditional credentials (Debroy & Auguste, 2025). As industries continue to evolve along with the advancement of technologies and globalization, there is a growing demand for skills such as digital literacy, interpersonal skills, and self-management skills (World Economic Forum, 2025). The future demands talents who can think critically, collaborate effectively, and continuously adapt to new environments and changes quickly with an open-mind. Programs like those developed by Opportunity@Work are necessary to respond to industrial demands.

All these examples show that CBE supports diverse learning pathways and acknowledges that learners acquire skills and knowledge through various experiences from both in-school and out-of-school settings. The shift of focusing

from time-based learning to mastery of skills requires a corresponding shift in how student progress and learning outcomes should be measured (OECD, 2018). Traditional assessments that focus on content knowledge and rote learning are insufficient for capturing the broader range of skills necessary for the future. Assessments must be transformed to evaluate not only what students know but also what they can do with that knowledge in real-world contexts (National Research Council, 2001). Similarly, the need for changes in admission and hiring systems is also becoming increasingly evident (Debroy & Auguste, 2025; Liu, 2021). Traditional systems that rely heavily on seat-time requirements of completing prerequisite courses may not fully capture a student's future readiness. Instead, demands of skills-based admissions and alternative credentialing models may be on the rise.

This chapter focuses on articulating design principles for educational assessments that address a skills focus to meet both academic and workforce needs. The discussion is situated in the context of the Skills for the Future (SFF) initiative (Liu et al., 2024; Ober et al., 2025b), a partnership between the ETS and Carnegie Foundation for the Advancement of Teaching. SFF serves three primary goals to measure what matters, develop innovative measures, and generate insights for key stakeholders. It aims to expand beyond traditional disciplinary learning by focusing on durable skills that matter in young learners' academic and workforce success. It also experiments on how student experience from a wide range of sources (e.g., school, family, community, workplace) can be considered to build a learner skills profile, through both innovative assessment and non-assessment evidential tools. Last, to address the information gaps in many previous assessments in which teachers and other stakeholders struggle to make sense out of the assessment results, SFF aims to adopt a co-design approach with educators and other stakeholders to best understand how assessment results can turn into insights for teaching and learning improvement.

The following chapter begins with a brief historical review of previous efforts at measuring a broader set of student skills. Then it reviews existing skills frameworks that outline key skills, competencies, and learning outcomes across various K–12, postsecondary, and workforce contexts. The review helps to identify gaps in existing frameworks and create a comprehensive taxonomy of skills for educational, occupational, and civic success, which will serve as a blueprint for future skills-based assessments being explored in SFF. The

chapter also discusses how technology and AI tools are used to capture skills acquisition, particularly the skills that are developed and demonstrated in non-traditional contexts. Then, the chapter discusses the assessment design principles that enable the measurement of complex skills with validity, reliability, and authenticity. Finally, the chapter proposes a professional development model and continuous improvement approach that supports the implementation of skills assessment in classrooms for SFF.

Previous Efforts to Measure a Broader Set of Student Skills

The past twenty years have seen an increasing and enduring interest in measuring a broader set of student skills beyond traditional academics. Many terms have been used to describe non-disciplinary skills such as 21st century skills, durable skills, transferable skills, employability skills, and the like (Trilling & Fadel, 2009). There is also considerable variation with regard to how frameworks define specific skills, provide guidance for possible assessments, and offer contexts of administration and use.

The Partnership for 21st Century Learning (P21) is one of the earliest collaborative initiatives seeking to infuse 21st century skills into education (Battelle for Kids, 2019). It defines key skills such as critical thinking, communication, collaboration, and creativity, and offers frameworks for educators to integrate these skills into curricula. P21 provides tools, resources, and professional development to a broad partnership of educators. While primarily focused on foundational issues such as identification and definition of the relevant skills, P21 also identified the need for and offered prototypes of associated assessments.

The Cognitive Readiness (CR) initiative of the US Department of Defense has made substantial investment in assessments of skills and traits closely related to the 21st century skills (Morrison & Fletcher, 2001). CR focuses primarily on human decision making in complex and stressful situations, endeavoring to develop the preconditions and skills necessary for effective decision making in military contexts. They employ innovative technologies such as simulations through virtual reality to design assessments for the targeted skills.

Assessing and Teaching 21st Century Skills (ATC21S) is a research initiative that aims to develop assessments for 21st century skills (Griffin et al., 2012). It focuses on defining, assessing, and integrating skills like collaboration, critical thinking,

and communication into educational frameworks. ATC21S has produced a set of innovative assessment tools for educators to evaluate students' 21st century skills. It has involved collaboration across a number of countries, leading to a rich exchange of ideas and practices. The project has generated substantial research on how to effectively assess these skills, contributing both specific tools and broader understanding of how to develop them.

The above-mentioned work, along with others (e.g., Pellegrino & Hilton, 2012; Cavanagh, 2010), provide early evidence for: (1) Demonstrations of framework development, dissemination, and adoption, and (2) Prototyping, testing, and refining approaches to assessing these nontraditional skills. These initiatives also helped promote awareness, understanding, and appreciation of the importance of the 21st century skills for learning and life.

At the same time, prior skills efforts also revealed challenges in measuring new, nontraditional skills in the complex contexts of the real world, data privacy concerns, integrating new forms of assessment into existing instructional and learning activity sets, professional development for educators' successful implementation, and assessment scalability in diverse educational settings.

SFF aims to draw upon previous efforts in executing its three goals in expanding what to measure, innovating how to measure, and generating insights. The following section discusses in detail prior assessment frameworks for complex skills, and describes a skills taxonomy that guides the assessment development for SFF.

Skills that Matter: A Review of Existing Skills-Based Educational Efforts

To more deeply understand the landscape of skill-based education systems, we conducted a review of current initiatives focusing on defining and assessing competencies across K–12, postsecondary, and workforce sectors. Our review included various skills frameworks and states' Portrait of a Graduate initiatives to identify priority skills of shared interests. Across major skills frameworks (The Collaborative for Academic, Social, and Emotional Learning [CASEL, 2020]; XQ student performance framework [XQ Institute, 2023]; OECD Learning Compass 2030 [OECD, n.d.]; NGLC MyWays Student Success Framework [Lash & Belfiore, 2017]; the European Framework for Personal, Social and Learning to Learn Key Competence [Sala et al., 2020]; Habits of Mind: 16 Essential Characteristics for Success [Institute for Habits of Mind, n.d.]; and Asia Society /CCSSO Global

Competence [Asia Society, 2013]), there is a significant overlap in social-emotional skills such as self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. Overlaps in these skills highlight the crucial roles these skills play in navigating complex, interconnected, and globalized worlds (Kim, Allen, & Jimerson, 2024). In addition, there is a strong emphasis on 21st century skills such as collaboration, communication, critical thinking, problem-solving, and creativity, which highlights the shift in educational goals toward preparing learners for the demands of the future workforce (Burning Glass Institute, 2023; National Research Council, 2012; Liu et al., 2023). These 21st century skills are becoming increasingly important as routine, repetitive tasks are being rapidly automated and unique human expertise plays a defining role in individuals' career success. Digital literacy and adaptability are especially emphasized in workforce-aligned frameworks (World Economic Forum, 2025; Burning Glass Institute, 2023), which also reflects the changing nature of future work and life driven by rapid technological advancements and industrial evolutions.

In analyzing these frameworks, it became evident that there was a need for clearer, more concrete definitions for many of the frequently cited skills. A notable pattern across the frameworks was the varying grainsize when skills are defined. Skills defined at broad levels often lack explicit definitions, making it difficult to understand the dimensions and sub-dimensions that underly the skills. For example, self-awareness is categorized as a broad competency with nine subskills in the CASEL framework (2020). In contrast, in the XQ framework (2023), self-awareness is positioned as a specific competency within the broader category of Learners for Life. This variation across frameworks illustrates how the same skill can be interpreted very differently, leading to potential confusion for educators attempting to implement these models.

Portrait of a Graduate (PoG) frameworks have also gained popularity in the United States. These frameworks are developed by individual states, outlining key competencies expected of their high school graduates. As of 2025, over 40 U.S. states have developed or are in the process of developing a PoG framework (Howard Terrell et al., 2025). We reviewed the PoG frameworks from 22 states which have provided adequate competency definitions. Several key skills emerged as common priorities across the majority of states (See Table 1). Communication was the most frequently mentioned skill, appearing in 21 out of 22 frameworks. Critical thinking and problem solving followed closely, mentioned by 19 and 17

states respectively. Collaboration was cited by 17 states. Other notable skills include civic engagement (13), perseverance (9), creativity (7), and growth mindset (7). The overlaps in essential skills across states suggest a shared vision for preparing K–12 students with a blend of cognitive, interpersonal, and personal competencies. The focus on these shared priority skills aligns with the demands of the 21st century workforce.

Table 1.
Overlaps in Skills Mentioned in States' PoG Frameworks.

Skill	# of States mentioned
Communication	21
Critical Thinking	19
Problem Solving	17
Collaboration	17
Civic Engagement	13
Perseverance	9
Growth Mindset	7
Creativity	8
Digital Literacy	7

A Comprehensive Taxonomy for the Skills for the Future

Creation of the Skills for the Future Taxonomy

The authors, along with a broader ETS research team, reviewed the broad and specific dimensions featured in all of the skills taxonomies and examined consistencies and discrepancies across the frameworks in terms of the names and definitions of dimensions. Through an iterative, consensus-seeking discussion, they then derived 30 "meta-dimensions" that cut across many of the frameworks. These dimensions form the basis of the integrative and comprehensive framework for SFF. A synthetic definition is provided for each meta-dimension, drawing on those revealed in the frameworks that were reviewed[2] (Table 2).

2 As with any term traceable to everyday speech (Cartwright & Bradburn, 2011), various sources—including frameworks we reviewed–define competencies and skills in different ways (e.g., Levine, 2021; Martinaitis, 2014; OECD, 2018; Soto et al., 2021). For our purposes we define a skill or competency as "a learned ability to perform an activity well".

Table 2.
Skills for the Future Taxonomy

Name	Major Skills
Adaptability	Working effectively in uncertain situations with shifting priorities by modifying one's actions or learning new skills in light of changing tasks and goals
Building Relationships	Understanding the importance of trust, respect for human dignity, and equality, and using these principles to establish and maintain healthy and supportive relationships, negotiate conflict constructively, and navigate interactions with diverse individuals and groups
Civic Engagement	Playing an active role in the global and local community and the application of civic values
Collaboration	Working with others cooperatively and coordinating effectively to achieve collective goals
Communication	Use of context-relevant strategies, domain-specific codes and tools when interacting with others, including active listening, asking questions, synthesizing messages, storytelling, and public speaking
Compassion	Feeling of sympathy with another person's feelings of sorrow or distress, often involving a desire to help or comfort that person
Creativity	Production or development of novel and useful outputs (e.g., understanding, perspectives, ideas, theories, products)
Critical Thinking	Understanding, managing, and analyzing information and arguments by making sound inferences, recognizing and evaluating assumptions, seeing rational connections, identifying patterns, constructing knowledge, and drawing evidence-based conclusions
Curiosity	The drive to investigate novel stimuli, including situations, people, and bodies of knowledge
Decision-Making	The cognitive processes and actions that result in choosing between two or more alternatives.
Digital Literacy	Creating, consuming, analyzing, and adapting in productive and responsible ways to utilize technology and communication tools in social, academic, and professional settings

Table 2. (continued)

Name	Major Skills
Disciplinary Literacies	Academic or subject specific literacy enabling learners to read, write, and speak like experts in a particular subject, including disciplinary knowledge, practices, and application skills
Educational & Occupational Awareness	Perception or knowledge of environments, people, facts, principles, and rules concerning school- or work-related topics and settings
Educational & Occupational Attitudes	Relatively enduring and general evaluations of objects relevant to school or work that exist on an emotional dimension ranging from negative to positive that influence one's approach to ideas, persons, and situations associated with educational or occupational settings
Educational & Occupational Values	Internal representations and perceptions of who one is as a person and how one wishes to define and lead a meaningful and satisfying life through their educational and occupational careers
Empathy	Vicarious experience of another person's feelings, emotions, and perspectives.
Growth Mindset	The belief that talents can be developed through persistent work, learning from risk taking and mistakes, and input from others
Leadership	Processes involved in directing others' efforts toward achieving individual, group, and/or organizational goals
Lifelong Learning	Understanding that learning takes place across the lifespan, having a positive attitude toward acquiring new skills across the lifespan, and engaging in acquiring new skills across the lifespan
Metacognition	Thinking about one's own cognition
People Skills	Behavioral interactions and behaviors to understand and manage the feelings of other individuals in team and other group settings to achieve individual or collective goals and develop productive working relationship to minimize conflict and maximize rapport
Perseverance	Overcoming obstacles and challenges by maintaining focus in the face of negative emotions, pursuing alternative routes to goal achievement, and persisting until the task is completed

Table 2. (continued)

Name	Major Skills
Problem Solving	The mental processes individuals use when they formulate plans and translate them into prospective actions for identifying a problem, gathering and evaluating information, developing solution paths, executing action plans, attempting to overcome difficulties, drawing conclusions, and adjusting to situational changes
Reasoning	Logic-based thinking processes of an inductive or deductive nature that are used to draw evidence-based conclusions from data, facts, or premises
Systems Thinking	Mental analyses of any system in order to understand system elements, the interconnections among the elements that drive the system to work as a whole, and how its constituent elements function both individually and in relation to each other
Self-Regulation	Regulating one's cognition and affect across different situations to maintain high motivation and energy through pursuing one's goals and restorative activities
Sensemaking	Gathering and interpreting data to rationalize and understand personal experiences and the world they live in and develop a personal sense of meaning
Stress Management	Regulating and decreasing stress via behavioral activities (e.g., breathing techniques, meditation) to stay positive, practice gratitude, and find ways to let go of worry
Taking Initiative	Proactively taking the first step in a task, enterprise, or process
Transformative Competencies	Competencies to transform the society and shape one's future to address the growing need to be innovative, responsible, and aware, including abilities to create new value, resolving and reconciling tensions and dilemmas, and taking responsibility

The SFF skills taxonomy consists of three primary domains (Danziger, 1994; Wilt & Revelle, 2019): affect (what & how people feel), behavior (what people do & how they do it), and cognition (what & how people think). The K–12 system explicitly rewards students' achievement in the cognitive domain by awarding high grades for the demonstration of knowledge in specific courses. While academic achievement may be facilitated by demonstrating some affective and behavioral skills (e.g., collaborating with other students to study effectively, remaining calm when taking challenging exams), those skills are simply a means to an end and not in and of themselves recognized as valuable by current K–12 structures. Aligned with many prominent frameworks, the SFF skills taxonomy emphasizes competencies beyond those represented by academic achievement for learners' future educational and occupational success.

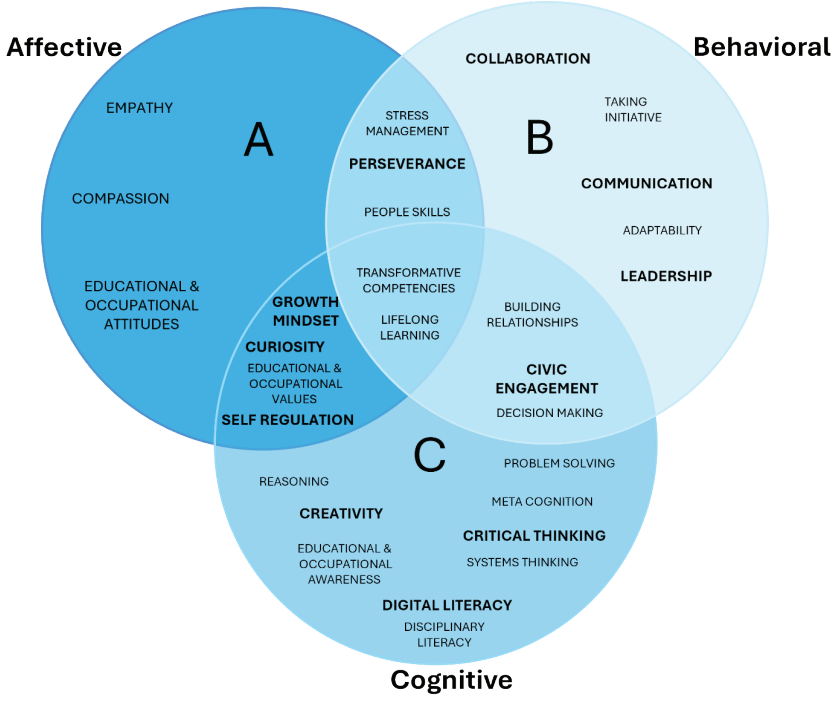
The research team independently classified the skills in the SFF taxonomy according to whether they best belonged to the affective, behavioral, or cognitive domains, based on the content of the competencies' definitions. Initial agreement among the team members was 86% for skills assigned to the affective category, 84% for behavioral skills, and 81% for cognitive skills. The researchers then met to resolve discrepancies in their classifications and collectively identified a category for the skills classification. The final results of the classifications are shown in Figure 1.

It is noteworthy that a subset of the competencies in the SFF Framework belong to more than one psychological domain. Each skill was initially assigned to a single domain that it was mostly aligned with. Through addressing the coding discrepancies in skill assignment, it became clear that some of the discrepancies stemmed from the fact that some skills fall into multiple categories. For example, *Building Relationships* is clearly behavioral in nature as its definition heavily relies on actions directed toward other human beings (e.g., navigating interactions, negotiating conflict). Yet, the definition also specifies that these actions are underwritten by cognitive understanding of various principles (e.g., equality, respect for human dignity), leading to the conclusion that it is more appropriate to classify Building Relationships as both a behavioral and cognitive skill. By the same token, the definition of *Lifelong Learning* contains elements that are affective (e.g., positive attitude toward learning), behavioral (e.g., acting to acquire new skills), and cognitive (e.g., understanding that learning can occur throughout life), suggesting that sorting it into a single domain would fail to capture its full breadth and complexity. Assigning the skills to multiple domains reflects the richness and complexity of these skills.

Figure 1. SFF Skills Taxonomy.

SFF Components Classified According to Affective, Behavioral, and Cognitive Domains

Note. The skills that are bolded represent those that were most prominent in our review of existing skills frameworks.



Why do Skills for the Future Matter?

Skills captured in the SFF taxonomy predict important education, career, and life outcomes. Affective and behavioral skills tend to predict the same outcomes as cognitive skills—and often with a similar degree of accuracy (Roberts et al., 2017). Although evidence for the practical importance of affective and behavioral skills has been accumulating since at least the 1970s (Bowles & Gintis, 1977; Jencks, 1979), they remain underemphasized in K–12 settings. This is particularly unfortunate given the many valuable life outcomes these types of skills have been consistently found to predict. Perseverance, for example, is related to educational attainment (Zamarro et al., 2018), salary (Ng et al., 2005), and longevity (Kern & Friedman, 2008), while empathy is associated with job performance (Sackett et al., 2022), civic participation (Ackermann, 2019), and health (Strickhouser et al., 2017). Many of these affective and behavioral skills are powerful predictors on their own, with their ability to forecast important outcomes only growing when they are considered in tandem (e.g., Ahadi & Diener, 1989).

Table 3.

Real-World Outcomes Predicted by Affective, Behavioral, and Cognitive Skills

Outcome	Predicted by Affective Skills	Predicted by Behavioral Skills	Predicted by Cognitive Skills
Educational	**Educational attainment** (Hampson et al., 2007) **K–12 grades** (Poropat, 2009) **Postsecondary grades** (Richardson et al., 2012)	**Educational attainment** (Zamarro et al., 2018) **K–12 grades** (Poropat, 2009) **Postsecondary grades** (Richardson et al., 2012)	**Educational attainment** (Brown et al., 2021) **K–12 grades** (Galla et al., 2019) **Postsecondary grades** (Richardson et al., 2012)

Table 3. (continued)

Outcome	Predicted by Affective Skills	Predicted by Behavioral Skills	Predicted by Cognitive Skills
Occupational	**Career choice** (Ackerman & Beier, 2003) **Career satisfaction** (Ng et al., 2005) **Job performance** (Sackett et al., 2022) **Job satisfaction** (Judge et al., 2002) **Salary** (Ng et al., 2005)	**Career choice** (Ackerman & Beier, 2003) **Career satisfaction** (Ng et al., 2005) **Job performance** (Connelly & Ones, 2010) **Job satisfaction** (Judge et al., 2002) **Salary** (Ng et al., 2005)	**Career choice** (Wai et al., 2009) **Grant funding** (Bernstein et al., 2019) h-index (Bernstein et al., 2019) **Income/salary** (Ng et al., 2005) **Job performance** (Nye et al., 2022) **Job prestige** (Lang & Kell, 2020) **Scholarly productivity** (Kuncel & Hezlett, 2007)
Civic	**Volunteerism** (McCann, 2017) **Voting** (Obschonka et al., 2018)	**Volunteerism** (Ackermann, 2019) **Voting** (Bakker et al., 2016)	**Volunteerism** (Proulx et al., 2018) **Voting** (Deary et al., 2008)
Health	**Longevity** (Friedman et al., 2010) **Mental health** (Strickhouser et al., 2017) **Physical health** (Rochefort et al., 2019)	**Longevity** (Kern et al., 2014) **Mental health** (Strickhouser et al., 2017) **Physical health** (Hampson et al., 2013)	**Longevity** (Calvin et al., 2011) **Mental health** (Davis & Humphrey, 2012) **Physical health** (Judge et al., 2010)

Are Skills for the Future Malleable?

Contemporary research shows that cognitive skills can be improved via participation in educational systems (Carlsson et al., 2015; Lehman et al., 1988; Ritchie et al., 2015; Ritchie & Tucker-Drob, 2018; Tock & Ericsson, 2019) and targeted interventions (Humphreys et al., 2022; Protzko, 2017; Protzko et al., 2013). Similarly, comprehensive meta-analyses of affective and behavioral skill interventions implemented among K–12 students (Cipriano et al., 2023; Durlak et al., 2011; Taylor et al., 2017) consistently show those interventions to be effective. Affective and behavioral skills have also been shown to be malleable via purposeful intervention in workforce, clinical, and community settings (Bleidorn et al., 2019; Martín-Raugh et al., 2022). Effective avenues for intervention include clinical treatment (Roberts et al., 2017), cognitive-behavioral therapy (Vittengl et al., 2004), social skills training (Piedmont, 2001), cognitive intervention (Jackson et al., 2012), mindfulness training (Krasner et al., 2009), situational judgment tests (Barron et al., 2022), developing and following developmental plans (Hudson et al., 2019), team-based training (Salas et al., 2008), coaching (Jones et al., 2016), and digital interventions (Allemand et al., 2023; Stieger et al., 2021).

Design Principles for Educational Assessment: Measuring Skills for the Future

There have been many efforts to incorporate non-academic skills in K–12 education. For example, 49 U.S. states and the District of Columbia have at least one policy that supports social-emotional learning (SEL) in schools, and 83% of U.S. school principals reported adopting a SEL curriculum (Skoog-Hoffman et al., 2024). Despite that many schools implement SEL, very few report scores on these skills, due to concerns about privacy, validity of assessment tools, and misuse of data (Skoog-Hoffman et al., 2024). Given the need for students to demonstrate a broader set of skills, approaches to help quantify learners' mastery of these skills are urgently needed.

A comprehensive assessment system is essential to provide a fuller understanding of what students can do and to guide their future learning pathways (Woo & Diliberti, 2022). This system must be rooted in rigorous research and innovation, featuring refined and new constructs, innovative task designs, breakthrough measurement sciences, advancements in measurement science (Wilson et al., 2005), sophisticated psychometric modeling (Embretson & Reise, 2013), precise

and reliable scoring methods (both human and automated; Bennett & Zhang, 2015), and accessible and actionable score reporting (Brookhart, 2013). SFF reimagines a skills-based assessment system with the following principles.

Five Assessment Principles

The SFF assessment system will encompass innovative assessments, an insights system that benefits multiple stakeholders including learners, educators, districts and states, and a professional learning community for educators. The skills featured in the system will be clearly and operationally defined, with SFF assessment development guided by five authentic assessment principles (McArthur, 2023; Palm, 2008; Sokhanvar et al., 2021).

Principle One: Reflecting the social and cultural backgrounds of students.

Students bring rich social, cultural, and linguistic backgrounds to the assessment experience (Elwood & Murphy, 2015). Assessments must fully embrace the diverse social and cultural backgrounds of the people who will be taking them (Lane, 2020). This requires the integration of culturally responsive assessment design, which considers linguistic diversity, varied ways of knowing, and equitable access to content and format (Gay, 2018). The SFF assessment system aims to bridge the gap between traditional assessments and the real-world applications of skills by incorporating authentic, context-rich tasks that mirror real-life and workplace experiences (Pellegrino & Hilton, 2012). By embedding tasks in meaningful and engaging scenarios, the system allows learners to demonstrate their competencies in ways that align with their lived experiences, ensuring a more holistic and equitable measurement of their abilities (Darling-Hammond et al., 2013). This approach not only enhances motivation and relevance for diverse learners but also improves the validity of assessment outcomes, as it captures a more comprehensive picture of their skills while minimizing cultural and contextual biases (Mislevy, 2018).

Principle Two: Centering around equity and fairness.

Persistent ethnic and racial performance differences in academic achievement have long been a critical concern in the United States, reflecting systemic inequities in educational opportunities, resources, and access to high-quality instruction (Ladson-Billings, 2006). In 2019, only 21% of all 12th-grade students demonstrated proficiency in mathematics, with significantly lower rates among historically

marginalized groups—just 11% of Latina/o/x students and 7% of African American students—highlighting enduring disparities in STEM education (United States Census Bureau, 2021; National Center for Education Statistics [NCES], 2020). The SFF assessment focuses on capturing a broad range of skills and knowledge in ways that reflect the varied experiences and strengths of learners, rather than favoring those who have had access to more traditional forms of academic preparation. The next generation of assessments must be designed to provide meaningful opportunities for all learners to demonstrate their abilities, serving as a tool for expanding access to educational and career pathways (Darling-Hammond et al., 2014). By incorporating real-world tasks, leveraging flexible assessment formats, and ensuring that measures are adaptable to different learning backgrounds, the SFF assessment aims to create a more effective and accurate representation of individuals' capabilities, ultimately helping to remove unnecessary barriers to success (Ober et al., 2025a; Liu et al., in press).

Principle Three: Benefiting instruction and learning.
The SFF assessment captures a broad spectrum of learners' abilities, going beyond traditional right-or-wrong scoring models to measure complex cognitive, affective, and behavioral skills. For example, when gathering evidence of students' critical thinking skills, the SFF assessment includes both direct assessment of students' critical thinking but also educators' submission of authentic evidence of students' critical thinking. By analyzing rich performance data, including students' problem-solving processes, decision-making strategies, and collaborative interactions, the system will generate actionable insights that can guide both individualized learning pathways and system-wide instructional improvements. These insights, provided at both the individual and cohort levels, aim to help students increase awareness of the skills that matter and understand their own skills level, and to help educators to incorporate skills in disciplinary instruction.

Principle Four: Using technology responsibly to generate insights.
The SFF system will leverage technological advancements in automated scoring and AI-supported assessments that are purposefully designed to enhance learning rather than simply introduce new tools without meaningful impact (Williamson et al., 2020). Beyond scoring, AI can support assessment design by analyzing large-scale learning data to identify key skill gaps, ensuring that assessments are aligned with real-world competencies and personalized learning needs (Mislevy, 2018).

When used responsibly, AI does not replace human judgment but rather augments educators' expertise by automating repetitive tasks, generating real-time feedback, and informing curriculum improvements, ultimately allowing teachers to focus on engaging students in deeper learning experiences (Dede, 2019).

Principle Five: Enabling personalization.
The SFF assessments will incorporate personalized choices for students to select the skills they want to be assessed about, the context of the skills, and ways of evidence demonstration. Personalized assessments allow learners to engage with tasks in ways that align with their unique strengths and learning pathways, leading to richer and more accurate insights about their abilities (Shute & Rahimi, 2021; Mislevy, 2018). Its insights reports aim to provide actionable, real-time feedback, offering a holistic view of what learners know and can do, as well as guidance on how to interpret and apply these insights for educational and career decision-making. These reports will be dynamic, diagnostic, and continuous, evolving with the learner to track progress over time rather than offering a single snapshot of performance (Bennett, 2018). By integrating real-time analytics, AI-driven feedback, and predictive modeling, assessment systems can support informed decision-making in areas such as admissions, educational progression, and workforce hiring (Williamson et al., 2020). Ultimately, this transformation in assessment design aims to empower learners, educators, and employers with deeper, more actionable insights that enable ongoing learning and skill development (Zieky & Perie, 2021).

Measuring Complex Skills Through Multimodal Assessment
SFF assessment will incorporate multimodal formats to enable learners to demonstrate their skills through diverse modalities, such as speech, gestures, writing, and digital interactions (Jaques et al., 2021). Multimodal assessment moves beyond traditional text-based responses, allowing for more authentic, interactive, and adaptive demonstrations of skills (Shute & Rahimi, 2021). Multimodal approaches expand the dimensions of skills that assessments can accurately capture, enabling learners to showcase what they can do in ways unattainable through traditional, single-modality assessment (e.g., reading, writing).

For example, traditionally oral communication is assessed in terms of aspects of verbal utterances, such as word choice, grammar, sentence structure, and tone. Multimodal assessment goes beyond this, uniting sensing technologies and machine learning to integrate information about nonverbal aspects of

communication, such as hand gestures, body posture, and facial expressions, leading to a more complete portrait of learners' skill in both the linguistic and social aspects of oral communication (Suendermann-Oeft et al., 2017). In the current digital age, holistic evaluations of students' learning are necessary to inform students of their achievements and needs as comprehensively as possible (Ross et al., 2020). By integrating information across multiple sensory modes (e.g., auditory, visual, written), multimodal assessment is perfectly poised to provide these holistic insights.

Advancements in multimodal technology allow greater insights into learners' skills. Multimodal assessment has been applied to a variety of domains including learners' English language proficiency (Forsyth et al., 2019), literacy (Tan et al., 2020), and collaborative learning and behavior (Khan, 2017). Relevant to multimodal assessment, multimodal analytics refers, as an example, to the inclusion of "advanced sensor technologies and machine learning systems to track and understand human behaviors" (Khan, 2017, p.175). Inferences from multiple sensory data can be made to draw conclusions about learners' proficiencies, abilities, attitudes, and dispositions.

Innovative Task Design

Accurately capturing SFF requires innovative task design. New assessment activities will go beyond traditional multiple-choice and constructed-response questions to enable the assessment of deep knowledge and thinking, reveal rich information about learners' interactions with the tasks (and, depending on the activity, other learners) through the generation of continuous process data, enable timely scoring at scale, and provide insights to help learners improve. Advancements in educational technology hold promise in enabling innovative task design. Immersive task environments can be designed to situate learners in authentic assessment situations. Game-based assessment offers simulation and interactivity, which expands the number and complexity of the constructs that can be measured precisely. The SFF system will use technology-rich environments to provide all learners with authenticity and interactivity during assessment experiences. In our application of advanced technological tools, we understand that digital tasks alone do not guarantee the quality of the assessment (Redecker & Johannessen, 2013). Research to date documents the value of a cognition-centered design approach to ensure the fidelity of the innovative tasks (Keehner, Arslan, & Lindner, 2023).

An illustrative example of what can be accomplished with cutting edge educational technology is the measurement of collaborative problem solving (CPS). CPS is a very complex construct that involves engaging with others in finding a solution to a commonly shared problem. Tasks that assess CPS well need to cover *both* collaboration and problem-solving dimensions. Once requiring grouping learners and closely observing their interactions, CPS appraisal can now be accomplished through interactive digital platforms that enable machine scoring at scale. ETS researchers have designed CPS tasks that leverage AI technology and data analytics (Hao, 2021; Hao et al., 2019). Collaboration and problem-solving skills are evaluated through authentic and virtual performance-based tasks. These tasks engage multiple learners simultaneously to solve a problem through an interactive assessment platform. The platform documents how individual learners share information, defend their stances, reconcile their opinions, and eventually identify a common solution. A chat function allows participants to display their problem-solving (cognitive) and collaborative (behavioral) skills dynamically as they interact with each other and the tasks themselves to come to solutions (Andrews-Todd & Forsyth, 2020).

Capturing Skills Gained from Multiple Pathways

An important goal of the SFF assessment is to recognize skills gained through alternative pathways, manifested in the K–12 to postsecondary transition, education to career transition, and occupation switch in the workforce. On the technological fronts, when inferences are made about individuals' skills through sources other than degrees and transcripts, evaluators often rely on self-report (e.g., cover letter, personal statement), third-party evaluation (e.g., reference letter, teacher rating)—data sources which have been found to highly favor wealthy students (Chetty et al., 2023)—or standardized assessment (e.g., cognitive test, personality inventory). New technology and widespread use of AI has enabled skills inference by parsing unstructured data (e.g., transcripts, resumes, employment history) into machine-readable data without the traditional evaluation (e.g., Sajjadiani et al., 2019). For example, teams at Experience You (T3 Innovation Network, n.d.), an initiative launched by the T3 Innovation Network and Education Design Lab, are working to turn unstructured data about individuals' educational, occupational, and experiential histories into quantitative, machine actionable data for documenting individuals' skills. The technologies and insights gained from these workforce initiatives hold great promise for high schools to offer credit for student learning that takes place outside of school, to overcome the barrier that information about

such activities (e.g., volunteering, internships, community service) is often available only in unstructured formats.

Broad considerations to equity issues should be embedded throughout the design, development, validation, and refinement of skills recognition and verification (Wilson & Martin, 2020). For example, when designing an analytical framework to capture skills from out of school experiences, it is important not to focus on extracurricular activities that are often only available to students from resourceful families. Playing piano, practicing swimming, and participating in a toastmaster program helps build resilience, perseverance, communication, and leadership skills. However, the SFF skills framework we apply to look for such skills should not just focus on these activities, as these activities may not be available to students from underprivileged backgrounds (Putnam, 2015). Equal consideration should be given to unstructured activities such as taking care of younger siblings, working at a local community shop, or even walking a far distance to school and being on time, as these activities represent resilience, perseverance, communication, and leadership as well (Larson, 2000). When conceptual framework and technological tools are used to capture skills, they need to be responsive to the experiences of students from all backgrounds.

What Types of Educational Experiences Promote the Development of Skills for the Future?

While the skills system we describe in the paper will provide valuable infrastructure and insights, for students to develop these skills, they will need access to new educational experiences. Traditional, didactic learning experiences in which students are asked to take in and regurgitate static information will not promote the development of skills for the future. Instead, through SFF assessment we hope to provide opportunities for students to demonstrate skills they gain from multiple authentic experiences, whether these experiences take place inside or outside of the schoolhouse and school day. By authentic, we mean learning experiences that are tied to actual performance and work associated with professions or academic disciplines (Collins & Duguid, 1989). In recent research, such experiences have been documented within extracurricular and elective experiences, where "Students were no longer vessels to be filled with knowledge, but rather people trying to produce something of real value," (Mehta & Fine, 2019). By Project-based, we mean learning experiences that are connected to real-world problems and contexts, driven by collaborative and social interactions among students, and students themselves

actively involved in the learning process (Kokotsaki, Menzies, & Wiggins, 2016). Teaching in authentic and project-based ways has been linked to the development of skills for the future such as collaboration, leadership, and communication (e.g., Vogler et al., 2018).

Supporting the Use of the SFF Assessment System

K–12 teachers will be critical to reimagining of the U.S. educational system through SFF. Incorporating SFF into teaching and learning and using the associated assessments effectively will require career-long development of ambitious pedagogy, including new instructional approaches that integrate SFF into disciplinary learning. Accordingly, teachers must be equipped with the instructional competencies, curricular materials, and assessment literacies to foster these skills within their students. For the SFF assessment system to be effectively executed in the classrooms, teachers' professional learning needs to be accompanied by strong communication and consistent engagement to develop buy-in with a wide range of stakeholders (e.g., parents, principals, superintendents).

Professional learning models to support SFF will necessitate implementation early in teachers' careers, including the pre-service and induction stages. To foster SFF affective and behavioral skills, in addition to cognitive competencies beyond Disciplinary Literacies, it will be essential for teachers to have strong content *and* pedagogical knowledge. Teachers proficient in both of these areas are more likely to organize high-quality curricula that engage students in complex problem solving (Hill et al., 2005) and teach in ways that help students construct, make meaning, evaluate, and test new knowledge (Cunningham, 1998; Windschitl et al., 2009). For professional learning ventures to be effective, they will have to imbue teachers with sophisticated reform-based practices (e.g., engaging in specialized discourses, relying on frequent assessment of student thinking, deep assessment literacy; Windschitl, 2009) needed to effectively nurture the integrated skill sets in students that are the defining feature of SFF. Key features of successful professional learning programs include sharing a vision for ambitious teaching and learning, relating teachers' learning to classroom practice, grounding the work in disciplinary teaching and learning, incorporating opportunities for active learning, and providing coherence with other learning activities (Darling-Hammond, 2000; Darling-Hammond et al., 2017; Garet et al., 2001). All of these elements, and more, will have to be developed to prepare teachers for educating students in SFF.

Improvement Science and Networks

The SFF system aims to provide insights report for educators to understand students' skill levels and provide guidance on skills improvement. However, providing the comprehensive support that teachers need to improve their practice based on these insights is not simple (Farrell & Marsh, 2016; Bertrand & Marsh, 2015). As has been documented in extensive research on educational program implementation, promoting improvements in practice at scale is beset with challenges (Honig, 2006). It's far easier to encourage the widespread adoption of shallow tweaks vs. deep change (McLaughlin & Mitra, 2001). The complexity of teaching means that "one size fits all" approaches to teacher learning are unlikely to lead to sustained improvements (Lampert, 2001). The political instability of educational organizations, such as districts, means that system leaders must be vigilant about creating and maintaining coherent instructional policies in order to encourage and sustain pedagogical improvement (Cobb et al., 2018). Furthermore, even when efforts at instructional improvement are able to overcome these challenges and demonstrate effectiveness in one location, they often struggle when brought to a new context (Coburn, 2003).

In response to these long-standing challenges of promoting wide-scale change, a new approach has gained popularity in education over the past decade: improvement science (Cohen-Vogel et al., 2015; Tichnor-Wagner et al., 2017). Improvement science is a systematic process of problem-solving that relies on the rapid refinement of innovations in response to data, a spirit of continuous inquiry, and sensitivity to local context (Langley et al., 2009). Rather than insisting on "fidelity" of implementation, it calls for the "adaptive integration" of new ideas into educational settings in such a way that honors the core design features of an innovation while simultaneously encouraging customization for local contexts (LeMahieu, 2011; Bryk et al., 2015). Practitioners of improvement science insist on the active incorporation of educators into the design, refinement, and execution of new practices.

Our approach to supporting educational organizations to use the SFF system will anchor itself in improvement science. Teachers and administrators are being involved in the co-design process for assessment development, as prior research (Windschitl et al., 2012) shows that educator involvement improves the alignment between assessment and instruction. Rather than being treated as passive recipients of "best practices," teachers are being involved in inquiry-based

professional communities that collectively examine assessment results, plan and implement changes to their practice, and use evidence to continuously refine their work. These communities are likely to provide collaborative and generative opportunities for teachers to understand the SFF framework and use it to decide how to connect the skills to curricula and instruction. Administrators, too, are currently taking part in inquiry groups that consider how to craft an inspiring instructional vision (Kay & Boss, 2021) and create policies that support the integration of these new assessments into their organization.

Alongside the use of improvement science principles, our approach to supporting educators will rely on the construction of learning networks that encourage the development of shared knowledge, the cross-pollination of ideas across educator groups, and the collective pursuit of improvement throughout a system (Russell et al., 2019). Rather than providing support to isolated schools or teacher teams, the SFF initiative brings together educators from various locations (schools within a district, or districts within a region), to work together to develop new ways to develop student skills. Recently, prominent philanthropies have invested heavily in the development of such improvement networks in the educational field (Bill & Melinda Gates Foundation, 2019). These networks can accelerate improvement by bringing together diverse sources of knowledge, energizing participants through productive collaboration, and providing a centralized source of learning (Kinlaw et al., 2020).

Conclusion

The current school assessment system limits its focus to a constrained set of knowledge and skills, typically easy to measure (Darling-Hammond et al., 2017; NRC, 2012; Schleicher, 2018; OECD, 2023). States have made attempts to support competence-based education out-of-schools (D'Brot, 2017), but assessment efforts to quantify learning gained from out of schools are limited. To prepare our next generation of learners for the challenges and opportunities of the future, a transformational assessment system is needed, one that is guided by sound assessment principles, captures learning acquired through multiple educational pathways, and offers ongoing and continuous insights for learners, teachers, post-secondary institutions, and employers. The assessment system SFF aims to design is guided by assessment principles, integrates measurement sciences, and offers personalized assessment to engage learners. New skills-based assessment

requires a paradigm shift from focusing on traditional cognitive skills to assessing and improving broader affective, behavioral, and cognitive skills that matter for life, work, and education. As a response to the paradigm shift, the SFF assessment system is a worthy experiment that builds on the already remarkable progress that has been made in competency-based and skills-based education in both K–12 and postsecondary education.

References

Ackerman, P. L., & Beier, M. E. (2003). Intelligence, personality, and interests in the career choice process. *Journal of Career Assessment, 11*(2), 205–218. https://doi.org/10.1177/1069072703011002006

Ackermann, K. (2019). Predisposed to volunteer? Personality traits and different forms of volunteering. *Nonprofit and Voluntary Sector Quarterly, 48*(6), 1119–1142. https://doi.org/10.1177/0899764019848484

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56*(3), 398–406. https://psycnet.apa.org/doi/10.1037/0022-3514.56.3.398

Allemand, M., Kirchberger, M., Milusheva, S., Newman, C., Roberts, B., & Thorne, V. (2023). *Conscientiousness and labor market returns*. World Bank. https://documents1.worldbank.org/curated/en/099355203272341682/pdf/IDU00a4b67200c3a70461a0b91b09fd8c4c97768.pdf

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior, 104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Asia Society. (2013). *Global Leadership*. Center for Global Education. https://asiasociety.org/sites/default/files/inline-files/all-grades-global-leadership-performance-outcomes-book-edu.pdf

Bakker, B. N., Rooduijn, M., & Schumacher, G. (2016). The psychological roots of populist voting: Evidence from the United States, the Netherlands and Germany. *European Journal of Political Research, 55*(2), 302–320. https://doi.org/10.1111/1475-6765.12121

Barron, L. G., Ogle, A. D., & Rowe, K. (2022). Improving the effectiveness of embedded behavioral health personnel through situational judgment training. *Military Psychology, 34*(4), 377–387. https://doi.org/10.1080/08995605.2021.1971938

Battelle for Kids, (2015). *Frameworks for 21st Century Learning*. http://www.battelleforkids.org/networks/p21

Beck, D., Morgado, L., & O'Shea, P. (2023). Educational practices and strategies with immersive learning environments: Mapping of reviews for using the metaverse. *IEEE Transactions on Learning Technologies*.

Bell, D. V. J. (2016). *Twenty-first century education: Transformative education for sustainability and responsible citizenship*. UNESCO/Brookings Institution. https://unesdoc.unesco.org/ark:/48223/pf0000245625

Bennett, R. E. (2018). Innovative assessment: The good, the bad, and the policy. *Education Inquiry, 9*(3), 299–317.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In *Technology and testing* (pp. 142–173). Routledge.

Bernstein, B. O., Lubinski, D., & Benbow, C. P. (2019). Psychological constellations assessed at age 13 predict distinct forms of eminence 35 years later. *Psychological Science, 30*(3), 444–454. https://doi.org/10.1177/0956797618822524

Bertrand, M., & Marsh, J. (2015). Teachers' Sensemaking of Data and Implications for Equity. *American Educational Research Journal, 52*(5), 861–893. https://doi.org/10.3102/0002831215599251

Bill & Melinda Gates Foundation. (2019). *Networks for School Improvement: Year One.* https://usprogram.gatesfoundation.org/What-We-Do/K-12-Education/Networks-for-School-Improvement

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining Twenty-First Century Skills. In P. Griffin, B. McGaw, & E. Care (Eds.) *Assessment and Teaching of 21st Century Skills.* (pp. 17–66). Dordrecht: Springer.

Bleidorn, W., Hill, P. L., Back, M. D., Denissen, J. J., Hennecke, M., Hopwood, C. J., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Wagner, J., Wrzus, C., Zimmermann, J., & Roberts, B. (2019). The policy relevance of personality traits. *American Psychologist, 74*(9), 1056. https://doi.org/10.1037/amp0000503

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD.

Bowles, S., & Gintis, H. (1977). *Schooling in capitalist America: Educational reform and the contradictions of economic life.* New York: Basic Books.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Brown, M. I., Wai, J., & Chabris, C. F. (2021). Can you ever be too smart for your own good? Comparing linear and nonlinear effects of cognitive ability on life outcomes. *Perspectives on Psychological Science*, *16*(6), 1337–1359. https://doi.org/10.1177/1745691620964122

Bryk, A., Gómez, L., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better.* Harvard Education Press.

Burning Glass Institute. (2023). *2023 Skills Compass Report*. https://www.burningglassinstitute.org/research/2023-skills-compass-report

Calvin, C. M., Deary, I. J., Fenton, C., Roberts, B. A., Der, G., Leckenby, N., & Batty, G. D. (2011). Intelligence in youth and all-cause-mortality: systematic review with meta-analysis. *International Journal of Epidemiology, 40*(3), 626–644. https://doi.org/10.1093/ije/dyq190

Care, E., Griffin, P., & Wilson, M. (Eds.). (2017). *Assessment and teaching of 21st century skills: Research and applications.* Springer.

Carlsson, M., Dahl, G. B., Öckert, B., & Rooth, D. O. (2015). The effect of schooling on cognitive skills. *Review of Economics and Statistics, 97*(3), 533–547. https://doi.org/10.1162/REST_a_00501

CASEL. (2020). *CASEL's SEL framework: What are the core competence areas and where are they promoted?* https://casel.org/casel-sel-framework-11-2020/?view=true

Cavanagh, S. (2010). *Common Core Standards: What They Mean for Education.* Education Week

Chamorro-Premuzic, T., & Frankiewicz, B. (2019, January 7). Does higher education still prepare people for jobs? *Harvard Business Review.* https://hbr.org/2019/01/does-higher-education-still-prepare-people-for-jobs

Chetty, R., Deming, D. J., & Friedman, J. N. (2023). *Diversifying society's leaders? The causal effects of admission to highly selective private colleges.* (No. w31492). National Bureau of Economic Research.

Cipriano, C., Strambler, M. J., Naples, L., Ha, C., Kirk, M. A., Wood, M., Sehgal, K., Zeiher, A., Eveleigh, A., McCarthy, M. F., Funaro, M., Ponnock, A., Chow, J., & Durlak, J. (2023). *Stage 2 report: The state of the evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions*. Child Development https://osf.io/mk35u/

Cobb, P., Jackson, K., Henrick, E., & Smith, T. M. (2018). *Systems for instructional improvement: Creating coherence from the classroom to the district office.* Harvard Education Press.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12. https://doi.org/10.3102/0013189X032006003

Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., & Wang, Q. (2015). Implementing educational innovations at scale: Transforming researchers into continuous improvement scientists. *Educational Policy, 29*(1), 257–277. https://doi.org/10.1177/0895904814560886

Collaborative for Academic, Social, and Emotional Learning. (2020). *CASEL'S SEL framework: What are the core competence areas and where are they promoted?* https://casel.org/casel-sel-framework-11-2020/

Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122. https://psycnet.apa.org/doi/10.1037/a0021212

Cunningham, C. M. (1998). The effect of teachers' sociological understanding of science (SUS) on curricular innovation. *Research in Science Education, 28*(2), 243–257. https://doi.org/10.1007/BF02462908

Danziger, K. (1994). *Constructing the subject: Historical origins of psychological research.* Cambridge University Press.

D'Brot, J. (2017). *Examining the validity structure of competency-based education.* U.S. Department of Education, Institute of Education Sciences, Regional Educational Laboratory Central. https://ies.ed.gov/sites/default/files/migrated/rel/regions/central/pdf/REL_2017249.pdf

Darling-Hammond, L. (2000). *Solving the dilemmas of teacher supply, demand and standards: How we can ensure a competent, caring, and qualified teacher for every child*. Columbia University, Teachers College, the National Commission on Teaching and America's Future.

Darling-Hammond, L., & Cook-Harvey, C. M. (2018). *Educating the whole child: Improving school climate to support student success*. Learning Policy Institute website: https://learningpolicyinstitute.org/media/547/download?inline&file=Educating_Whole_Child_REPORT.pdf

Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute website: https://learningpolicyinstitute.org/sites/default/files/product-files/Effective_Teacher_Professional_Development_REPORT.pdf

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives, 22*(86), 1–32.

Davis, S. K., & Humphrey, N. (2012). Emotional intelligence predicts adolescent mental health beyond personality and cognitive ability. *Personality and Individual Differences, 52*(2), 144–149. https://doi.org/10.1016/j.paid.2011.09.016

Deary, I. J., Batty, G. D., & Gale, C. R. (2008). Childhood intelligence predicts voter turnout, voting preferences, and political involvement in adulthood: The 1970 British Cohort Study. *Intelligence, 36*(6), 548–555. https://doi.org/10.1016/j.intell.2008.09.001

Debroy, P., & Auguste, B. (2025, July 8). Using AI to advance skills-first hiring. *Brookings Institution*. https://www.brookings.edu/articles/using-ai-to-advance-skills-first-hiring

Dede, C. (2019). *Artificial intelligence in education: Promise and implications for teaching and learning.* Harvard University Graduate School of Education.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education*. McMillan.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Elwood, J., & Murphy, P. (2015). Assessment systems as cultural scripts: A sociocultural theoretical lens on assessment practice and products. *Assessment in Education: Principles, Policy & Practice, 22*(2), 182–192.

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. Psychology Press.

Farrell, C. C., & Marsh, J. A. (2016). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education, 60*, 398–412. https://doi.org/10.1016/j.tate.2016.07.010

Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, G. T., Evanini, K., & So, Y. (2019). Evaluating English language learners' conversations: Man vs. Machine. *Computer Assisted Language Learning*, *32*(4), 398–417. https://doi.org/10.1080/09588221.2018.1517126

Friedman, H. S., Kern, M. L., & Reynolds, C. A. (2010). Personality and health, subjective well-being, and longevity. *Journal of Personality, 78*(1), 179–216. https://doi.org/10.1111/j.1467-6494.2009.00613.x

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal, 56*(6), 2077–2115. https://doi.org/10.3102/0002831219843292

García-Chitiva, M. del P. (2024). The centrality of soft skills in higher education: Theory, methodology and practice. In M. Shelley & O. T. Ozturk (Eds.), *Proceedings of ICRES 2024: International Conference on Research in Education and Science*. https://files.eric.ed.gov/fulltext/ED673093.pdf

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice.* Teachers College Press.

Gómez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies, 16*(4), 500–515.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st Century Skills.* Dordrecht: Springer. https://doi.org/10.1007/978-94-007-2324-5

Gutiérrez, K., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19–25.

Hampson, S. E., Edmonds, G. W., Goldberg, L. R., Dubanoski, J. P., & Hillier, T. A. (2013). Childhood conscientiousness relates to objectively measured adult physical health four decades later. *Health Psychology, 32*(8), 925–928. https://doi.org/10.1037/a0031655

Hampson, S. E., Goldberg, L. R., Vogt, T. M., & Dubanoski, J. P. (2007). Mechanisms by which childhood personality traits influence adult health status: Educational attainment and healthy behaviors. *Health Psychology, 26*(1), 121–125. https://doi.org/10.1037/0278–6133.26.1.121

Hao, J. (2021). Beyond a single score: Scoring and reporting strategy for scalable assessments of collaborative problem solving [Paper presentation]. ITC 2021 Symposium, New Constructs for the New Economy. Virtual.

Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. A. (2019). Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving. *ETS Research Report Series ETS RR-19–41*. https://doi.org/10.1002/ets2.12276

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406. https://doi.org/10.3102/00028312042002371

Honig, M. I. (Ed.). (2006). *New directions in education policy implementation: Confronting complexity*. State University of New York Press.

Howard Terrell, J., Ahigian, R., Garvey, M., & Barzee, S. (2025). *So, you want to create a Portrait of a Graduate? Factors and considerations for the field*. U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED673579.pdf

Hudson, N. W., Briley, D. A., Chopik, W. J., & Derringer, J. (2019). You have to follow through: Attaining behavioral change goals predicts volitional personality change. *Journal of Personality and Social Psychology, 117*(4), 839–857. https://doi.org/10.1037/pspp0000221

Humphreys, K. L., King, L. S., Guyon-Harris, K. L., Sheridan, M. A., McLaughlin, K. A., Radulescu, A., Nelson, C. A., Fox, N. A., & Zeanah, C. H. (2022). Foster care leads to sustained cognitive gains following severe early deprivation. *Proceedings of the National Academy of Sciences, 119*(38), Article e2119318119. https://doi.org/10.1073/pnas.2119318119

The Institute for Habits of Mind. (n.d.). Habits of Mind Framework. Retried from https://www.habitsofmindinstitute.org/wp-content/uploads/2021/03/HOM-Table-Large-Attribution.pdf

Jackson, J. J., Hill, P. L., Payne, B. R., Roberts, B. W., & Stine-Morrow, E. A. (2012). Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults. *Psychology and Aging, 27*(2), 286–292. https://doi.org/10.1037/a0025918

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2021). Multimodal learning analytics: Towards an integrated approach for understanding learning. *Artificial Intelligence in Education, 31*(2), 203–218.

Jencks, C. (1979). *Who gets ahead? The determinants of economic success in America.* Basic Books.

Jones, R. J., Woods, S. A., & Guillaume, Y. R. (2016). The effectiveness of workplace coaching: A meta-analysis of learning and performance outcomes from coaching. *Journal of Occupational and Organizational Psychology, 89*(2), 249–277. https://doi.org/10.1111/joop.12119

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541. https://doi.org/10.1037/0021–9010.87.3.530

Judge, T. A., Ilies, R., & Dimotakis, N. (2010). Are health and happiness the product of wisdom? The relationship of general mental ability to educational and occupational attainment, health, and well-being. *Journal of Applied Psychology, 95*(3), 454–468. https://doi.org/10.1037/a0019084

Kay, K., & Boss, S. (2021). *Redefining student success: Building a new vision to transform leading, teaching, and learning.* Corwin Press.

Keehner, M., Arslan, B., & Lindner, M. A. (2023). Cognition-centered design principles for digital assessment tasks and items. In R. J. Tierney, F. Rivzi, K. Ercikan (Eds.), *International Encyclopedia of Education,* (4th ed., pp.171–184). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10025-9

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980. https://doi.org/10.3102/0034654315626800

Kern, M. L., Della Porta, S. S., & Friedman, H. S. (2014). Lifelong pathways to longevity: Personality, relationships, flourishing, and health. *Journal of Personality, 82*(6), 472–484. https://doi.org/10.1111/jopy.12062

Kern, M. L., & Friedman, H. S. (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology, 27*(5), 505–512. https://doi.org/10.1037/0278–6133.27.5.5

Khan, S. M. (2017). Multimodal behavioral analytics in intelligent learning and assessment systems. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 173–184). Switzerland: Springer. https://doi.org/10.1007/978-3-319-33261-1_11

Kim, E. K., Allen, J. P., & Jimerson, S. R. (2024). Supporting student social emotional learning and development. *School Psychology Review, 53*(3), 201–207. https://doi.org/10.1080/2372966X.2024.2346443 https://psycnet.apa.org/record/2024–89197–001

Kinlaw, A., Snyder, M., Chu, E., Lau, M., Lee, S., & Nagarajan, P. (2020). *Managing for change: Achieving systemic reform through the effective implementation of networks for school improvement*. Center for Public Research and Leadership, Columbia University. https://cprl.law.columbia.edu/sites/default/files/content/docs/ManagingforChangevF.pdf

Kokotsaki, D., Menzies, V., & Wiggins, A. (2016). Project-Based Learning: A Review of the Literature. Improving Schools, 19, 267-277.

Krasner, M. S., Epstein, R. M., Beckman, H., Suchman, A. L., Chapman, B., Mooney, C. J., & Quill, T. E. (2009). Association of an educational program in mindful communication with burnout, empathy, and attitudes among primary care physicians. *Jama, 302*(12), 1284–1293. https://doi.org/10.1001/jama.2009.1384

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3–12

Lampert, M. (2001). *Teaching Problems and the Problems of Teaching*. Yale University Press.

Lane, S. (2020). Fairness and validity in educational assessment. *Educational Measurement: Issues and Practice, 39*(2), 20–30.

Lang, J. W., & Kell, H. J. (2020). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology, 105*(9), 1047–1061. https://psycnet.apa.org/doi/10.1037/apl0000472

Langley, G. J., Moen, R., Nolan, K. M., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. John Wiley & Sons.

Larson, R. W. (2000). Toward a psychology of positive youth development. *American Psychologist, 55*(1), 170–183. https://doi.org/10.1037/0003–066X.55.1.170

Lash, D., & Belfiore, G. (2017). Visual Summary of the MyWays Student Success Series. https://s3.amazonaws.com/nglc/resource-files/MyWays_000VisualSummary.pdf

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43*(6), 431–442. https://psycnet.apa.org/doi/10.1037/0003–066X.43.6.431

LeMahieu, P. (2011). *What we need is more integrity (and less fidelity) of implementation.* Carnegie Commons. https://www.carnegiefoundation.org/blog/what-we-need-in-education-is-more-integrity-and-lessfidelity-of-implementation/

Lench, S., Fukuda, E., & Anderson, R. (2015). *Essential skills and dispositions: Developmental frameworks for collaboration, creativity, communication, and self-direction.* Center for Innovation in Education at the University of Kentucky, https://www.inflexion.org/essential-skills-and-dispositions-development-frameworks/

Levine, E. (2021). *Habits of Success: Helping Students Develop Essential Skills for Learning, Work, and Life*. Aurora Institute. https://files.eric.ed.gov/fulltext/ED618110.pdf.

Levine, E., & Patrick, S. (2019). What Is Competency-Based Education? An Updated Definition. *Aurora Institute*. https://files.eric.ed.gov/fulltext/ED604019.pdf

Liu, L., Courey, K. A., Kinsey, D. Ober, T. M., &Johnson, D. G. (in press). Navigating the digital horizon: A proposed framework and strategies for assessing digital literacy. *ETS Research Report*.

Liu, O. L. (2021). Five trends that are reshaping the course of American higher education. *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊, 2*(3), Article 1. https://doi.org/10.59863/NDBC2976

Liu, O. L., Kell, H., Williams, K., Ling, G., & Sanders, M. (2023). ETS skills taxonomy 2025. *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊, 4*(4), 1. https://doi.org/10.59863/NMIE9603

Liu, O. L., Rios, J. A., & Bailey, A. L. (2022). Multimodal assessments: Opportunities and challenges in measuring 21st century skills. *Educational Assessment, 27*(1), 1–15.

Liu, O. L., Wang, Y., Liu, L., & Ling, G. (2024). Skills for the Future: A New Vision for Skills-Based Assessment. Paper presented at the *Promoting Competence-Based Education: Competence Frameworks and Classroom Implementation* session, annual conference of National Council for Measurement in Education (NCME). Philadelphia, PA.

Martín-Raugh, M., Kell, H., Ling, G., Fishtein, D., & Yang, Z. (2022). Noncognitive skills and critical thinking predict undergraduate academic performance. *Assessment & Evaluation in Higher Education, 48*(3), 350–361. https://doi.org/10.1080/02602938.2022.2073964

McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. *Higher Education, 85*(1), 85–101. https://doi.org/10.1007/s10734-022-00822-y

McCann, S. J. (2017). Higher USA state resident neuroticism is associated with lower state volunteering rates. *Personality and Social Psychology Bulletin, 43*(12), 1659–1674. https://doi.org/10.1177/0146167217724802

McLaughlin, M., & Mitra, D. (2001). Theory-based change and change-based theory: Going deeper, going broader. *Journal of Educational Change, 2*(4), 301–323. https://doi.org/10.1023/A:1014616908334

Mehta, J., & Fine, S. (2019, Mar. 30). High school doesn't have to be boring. *The New York Times*. https://larrycuban.wordpress.com/2019/04/04/high-school-doesnt-have-to-be-boring-jal-mehta-and-sarah-fine/

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Montessori, M. (1948). *To educate the human potential*. Kalakshetra Publications.

Morrison, J. E., Fletcher, J. D (2001). *Cognitive Readiness*. Defense Technical Information Center. https://apps.dtic.mil/sti/citations/tr/ADA417618.

National Association of Colleges and Employers. (2019a). *Career readiness for the new college graduate: A definition and competencies*. https://www.naceweb.org/uploadedfiles/pages/knowledge/articles/career-readiness-fact-sheet-jan-2019.pdf

National Center for Education Statistics (NCES). (2020). *NAEP 2019 mathematics and reading assessments.* U.S. Department of Education.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* National Academy Press. https://doi.org/10.17226/10019

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* National Academies Press. https://doi.org/10.17226/13398

National School Boards Association. (2025, April). *Research: Soft skills matter.* https://www.nsba.org/resources/asbj/asbj-april-2025/april-2025-research-soft-skills-matter

Ng, T. W., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology, 58*(2), 367–408. https://doi.org/10.1111/j.1744-6570.2005.00515.x

Nye, C. D., Ma, J., & Wee, S. (2022). Cognitive ability and job performance: Meta-analytic evidence for the validity of narrow cognitive abilities. *Journal of Business and Psychology, 37*(6), 1119–1139. https://doi.org/10.1007/s10869-022-09796-1

Ober, T. M., Liu, L., Nitkin, D., & Liu, O. L. (2025b). Aligning models of competency-based education with skills for the future. *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*, *6*(3), 1–15.

Ober, T., Johnson, D., Liu, L., Kinsey, D., & Courey, K. (2025a). Communication as a Future Ready Skill: A Proposed Framework and Strategies for Assessment. *ETS Research Report Series*, *2025*(1).

Obschonka, M., Stuetzer, M., Rentfrow, P. J., Lee, N., Potter, J., & Gosling, S. D. (2018). Fear, populism, and the geopolitical landscape: The "sleeper effect" of neurotic personality traits on regional voting behavior in the 2016 Brexit and Trump elections. *Social Psychological and Personality Science, 9*(3), 285–298. https://doi.org/10.1177/1948550618755874

Organisation for Economic Co-operation and Development (OECD). (2018). *The future of education and skills: Education 2030*. OECD Publishing. https://www.oecd.org/education/2030-project/

Organisation for Economic Co-operation and Development (OECD). (2023). *Innovating assessments to measure and support complex skills*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/04/innovating-assessments-to-measure-and-support-complex-skills_b0255009/e5f3e341-en.pdf

Organisation for Economic Co-operation and Development (OECD). (n.d.). *The OECD Learning Compass 2030*. https://www.oecd.org/en/data/tools/oecd-learning-compass-2030.html

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research, and Evaluation, 13*, 4. 1–11. https://doi.org/10.7275/0qpc-ws45

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* National Academies Press.

Piedmont, R. L. (2001). Cracking the plaster cast: Big Five personality change during intensive outpatient counseling. *Journal of Research in Personality, 35*(4), 500–520. https://doi.org/10.1006/jrpe.2001.2326

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin & Review, 24*, 1022- 1031. https://doi.org/10.3758/s13423-016-1196-1

Protzko, J., Aronson, J., & Blair, C. (2013). How to make a young child smarter: Evidence from the database of raising intelligence. *Perspectives on Psychological Science, 8*(1), 25–40. https://doi.org/10.1177/1745691612462585

Proulx, C. M., Curl, A. L., & Ermer, A. E. (2018). Longitudinal associations between formal volunteering and cognitive functioning. *The Journals of Gerontology: Series B, 73*(3), 522–531. https://doi.org/10.1093/geronb/gbx110

Putnam, R. D. (2015). *Our kids: The American dream in crisis*. Simon & Schuster.

Redecker, C., & Johannessen, Ø. (2013). *Changing assessment—Towards a new assessment paradigm using ICT*. European Journal of Education, 48(1), 79–96. https://doi.org/10.1111/ejed.12018

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. https://doi.org/10.1037/a0026838

Ritchie, S. J., Bates, T. C., & Deary, I. J. (2015). Is education associated with improvements in general cognitive ability, or in specific skills? *Developmental Psychology, 51*(5), 573–582. https://doi.org/10.1037/a0038981

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*(8), 1358–1369. https://doi.org/10.1177/0956797618774253

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin, 143*(2), 117–141. https://doi.org/10.1037/bul0000088

Rochefort, C., Hoerger, M., Turiano, N. A., & Duberstein, P. (2019). Big Five personality and health in adults with and without cancer. *Journal of Health Psychology, 24*(11), 1494–1504. https://doi.org/10.1177/1359105317753714

Ross, J., Curwood, J. S., & Bell, A. (2020). A multimodal assessment framework for higher education. E-*learning and Digital Media*, *17*(4), 290–306. https://doi.org/10.1177/2042753020927201

Russell, J. L., Bryk, A. S., Peurach, D., Sherer, D., Khachatryan, E., LeMahieu, P. G., Sherer, J. Z., & Hannan, M. (2019). *The social structure of networked improvement communities: Cultivating the emergence of a scientific-professional learning community* [Paper presentation]. American Educational Research Association Annual Meeting, Toronto, ON.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology, 104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Sala, A., Punie, Y., Garkov, V., & Cabrera Giraldez, M. (2020). *LifeComp: The European Framework for Personal, Social and Learning to Learn Key Competence*. Publications Office of the European Union, Luxembourg. doi:10.2760/302967

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors, 50*(6), 903–933. https://doi.org/10.1518/001872008X375009

Schleicher, A. (2018). *World class: How to build a 21st-century school system*. OECD Publishing. https://doi.org/10.1787/9789264300002-en

Shute, V. J., & Rahimi, S. (2021). Review of modern psychometrics and automated scoring: Process, product, and potential. *Educational Psychologist, 56*(2), 67–88.

Silva, E., White, T., & Toch, T. (2015). *The Carnegie Unit: A Century-Old Standard in a Changing Education Landscape*. Carnegie Foundation for the Advancement of Teaching. https://www.luminafoundation.org/files/resources/carnegie-unit-report.pdf

Skoog-Hoffman, A., Miller, A. A., Plate, R. C., Meyers, D. C., Tucker, A. S., Meyers, G., Diliberti, M. K., Schwartz, H. L., Kuhfeld, M., & Jagers, R. J. (2024). *Social and emotional learning in U.S. schools: Findings from CASEL's nationwide policy scan and the American Teacher Panel and American School Leader Panel surveys* (RR-A1822–2). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1822-2.html

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation, 70*, Article 101030. https://doi.org/10.1016/j.stueduc.2021.101030

Southern New Hampshire University. (n.d.). *Community partnerships: Competency-based education.* https://www.snhu.edu/about-us/social-impact/community-partnerships

Stafford-Brizard, K. B. (2016). *Building blocks for learning: A framework for comprehensive student development*. Turnaround for Children. https://turnaroundusa.org/wp-content/uploads/2016/03/Turnaround-for-Children-Building-Blocks-for-Learningx-2.pdf

Stieger, M., Flückiger, C., Rüegger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2021). Changing personality traits with the help of a digital personality change intervention. *Proceedings of the National Academy of Sciences, 118*(8), Article e2017548118. https://doi.org/10.1073/pnas.2017548118

Strickhouser, J. E., Zell, E., & Krizan, Z. (2017). Does personality predict health and well-being? A metasynthesis. *Health Psychology, 36*(8), 797–810. https://doi.org/10.1037/hea0000475

Suendermann-Oeft, D., Ramanarayanan, V., Yu, Z., Qian, Y., Evanini, K., Lange, P., Wang, X., & Zechner, K. (2017). *A multimodal dialog system for language assessment: Current state and future directions*. (ETS Research Report Series No. RR-17–21). ETS. https://doi.org/10.1002/ets2.12149

T3 Innovation Network. (n.d.). *Experience You*. U.S. Chamber of Commerce Foundation. https://www.t3networkhub.org/experienceyou

Tan, L., Zammit, K., D'warte, J., & Gearside, A. (2020). Assessing multimodal literacies in practice: A critical review of its implementations in educational settings. *Language and Education*, *34*(2), 97–114. https://doi.org/10.1080/09500782.2019.1708926

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development, 88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine*, *140*, 227–235.

Tichnor-Wagner, A., Wachen, J., Cannata, M., & Cohen-Vogel, L. (2017). Continuous improvement in the public school context: Understanding how educators respond to plan-do-study-act cycles. *Journal of Educational Change, 18*(4), 465–494. https://doi.org/10.1007/s10833-017-9301-4

Tock, J. L., & Ericsson, K. A. (2019). Effects of curricular emphasis in college on the GRE and its impact on the gender gap in performance. *Contemporary Educational Psychology, 56*, 40–54. https://doi.org/10.1016/j.cedpsych.2018.11.003

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey-Bass.

Vittengl, J. R., Clark, L. A., & Jarrett, R. B. (2004). Improvement in social-interpersonal functioning after cognitive therapy for recurrent depression. *Psychological Medicine*, *33*(4), 643–658. https://doi.org/10.1017/S0033291703001478

Ulloa-Cazarez, R. (2021). Soft skills and online higher education. In M.-T. Lepeley, N. J. Beutell, N. Abarca, & N. Majluf (Eds.), *Soft skills for human centered management and global sustainability* (pp. 77–92). Routledge. https://doi.org/10.4324/9781003094463-6-9

United States Census Bureau. (2021). Educational attainment in the United States: 2019.

Vogler, J. S., Thompson, P., Davis, D. W., Mayfield, B. E., Finley, P. M., & Yasseri, D. (2018). *The hard work of soft skills: augmenting the project-based learning experience with interdisciplinary teamwork. Instructional Science, 46*(3), 457-488.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817–835. https://doi.org/10.1037/a0016127

Western Governors University. (2019). *What is competency-based education?* https://www.wgu.edu/about/story/cbe.html

Werquin, P. (2023). *Formal, non-formal, and informal learning: What are they, and how do they differ?* ERIC. https://files.eric.ed.gov/fulltext/ED626005.pdf

Williamson, B., Eynon, R., & Potter, J. (2020). Pandemic politics, pedagogies and practices: Digital technologies and AI in the COVID-19 crisis. *Learning, Media and Technology, 45*(2), 107–114.

Wilson, B., & Martin, N. (2020). Equity and quality in skills recognition: Challenges and opportunities in digital credentialing. *International Journal of Educational Technology in Higher Education, 17*(1), 1–15. https://doi.org/10.1186/s41239–020–00213–2

Wilson, M. R., Bertenthal, M. W., & Wilson, M. R. (2005). *Systems for state science assessment* (Vol. 248). National Academies Press.

Wilt, J., & Revelle, W. (2019). The Big Five, everyday contexts and activities, and affective experience. *Personality and Individual Differences, 136*(1)*, 140–147.* https://doi.org/10.1016/j.paid.2017.12.032

Windschitl, M. (2009, February). *Cultivating 21st century skills in science learners: How systems of teacher preparation and professional development will have to evolve.* Presentation given at the National Academies of Science Workshop on 21st Century Skills, Washington, DC.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education, 96*(5), 878–903. https://doi.org/10.1002/sce.21027

Woo, A., & Diliberti, M. (2022). *The role of benchmark assessments in coherent instructional systems*. Rand Corporation, https://www.rand.org/content/dam/rand/pubs/research_reports/RRA100/RRA134-19/RAND_RRA134-19.pdf

World Economic Forum. (2025). *The Future of Jobs Report 2025*. Geneva: World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2025/

XQ Institute. (2023). *XQ competency rubric*. https://admin.xqsuperschool.org/wp-content/uploads/2023/03/XQ_Competency_Rubric_V1.1.pdf

Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics, 72*, 51–60.

Zieky, M., & Perie, M. (2021). The future of assessment: Measuring what matters in education and the workplace. Educational Testing Service.

Centering the Voices of Assessment Users in the Advancement of Early Learning Measures

Emily C. Hanno, Elizabeth Mokyr Horner, Ximena A. Portilla, and JoAnn Hsueh

Acknowledgments: This chapter is based on research funded by the Gates Foundation.

Abstract

Increasingly young children spend time in formalized learning settings before they enter kindergarten. Although this period can be an impressive time of growth and development for young learners, early education practitioners and leaders often lack easy-to-use, reliable, and valid tools to inform their work. This chapter describes the Measures for Early Success Initiative aimed at developing novel child assessments that accurately capture what all young learners know and can. This chapter begins by introducing the ambitious vision motivating the Measures for Early Success Initiative, describing the goals and features of child assessment tools that are likely to be usable and useful across today's early childhood education landscape. Then it describes the Measures for Early Success Initiative's approach to working towards this vision through inclusive, iterative research and development cycles involving interdisciplinary assessment developer teams working in collaboration with communities across the United States. Initial learnings from this approach underscore the value of integrating user perspectives in the assessment design and development process to ensure tools can be used in the service of learning. In line with principles underlying this Handbook, the chapter highlights promising approaches to support engagement in assessment activities and allow respondents to draw upon their background knowledge and experiences.

Introduction

Early childhood education programs intended to care for and educate children before they enter kindergarten are a promising approach for fostering healthy development and supporting working families. These programs can offer young children complex, dynamic environments in which they can interact, explore, and develop new skills and abilities that prepare them for success in elementary school and beyond (Yoshikawa et al., 2013). The evidence base on early childhood education programs underscores their potential to positively impact children, families, and communities (McCoy et al., 2017), yet it also illuminates challenges of scaling high-quality early learning systems. Not all children have access to the sorts of high-quality early childhood programs thought to confer a developmental boost (Jones et al., 2020), and pre-K-related benefits to children's skills at kindergarten entry tend to disappear quickly during early elementary school (Abenavoli, 2019). Understanding and addressing these unsolved challenges relies on having comprehensive, accurate information on how children's development progresses over time.

Data from assessments that capture children's skills can inform the work of early learning systems, educators, and families in supporting young children's development, as well as identify programs, policies, and practices that allow children to reach their full potential (deMonsabert et al., 2021; Im, 2017). Yet, several key limitations of most existing tools can make it challenging to regularly gather reliable insights into young children's skills at scale. First, most child assessments for early learners focus on narrow sets of skills that are not consistently linked with longer-term indicators of success (McCormick & Mattera, 2022). Second, most tools have been developed and validated with homogenous study samples that are not representative of the children enrolled in public pre-K, which means they may not yield accurate insights about all children (Hsueh, 2021). Third, child assessment data are often burdensome to collect, analyze, and act on in real-world settings. These limitations mean that families, educators, and systems are unlikely to have accurate insights into the strengths and needs of all children, as well as early learning programs. Responding to data from these tools may ultimately exacerbate false narratives about specific subpopulations of children and widen gaps in children's early learning experiences.

This chapter describes recent efforts aimed at addressing these limitations by improving the measurement of young children's outcomes to better meet the needs of assessment users, defined broadly as children, families, educators, administrators, systems, and researchers. Specifically, it outlines the progression and approach of the Measures for Early Success Initiative (or Measures Initiative), a large-scale research and development (R&D) initiative involving collaboration between researchers, practitioners, product developers, and technologists to develop innovative, evidence-based direct child assessments that are usable in and useful for public pre-K settings across the United States.[1]

The Measures Initiative focuses on identifying new practitioner-friendly direct assessment approaches, using methods that collect information from children through standardized tasks or activities as opposed to from observations or work sampling approaches. Tools coming out of this initiative are intended to be used by practitioners to inform instructional decisions but also yield insights that can speak to broader questions about programs and policies. For example, a center-based educator might use the tools to understand children's progress toward early math standards, while program leaders may also use them to consider whether additional math supports are needed program-wide.

The first section of this chapter outlines opportunities for reimagined direct assessments in the areas of content, psychometrics, experience, usefulness, and scalability that serve as the foundation for this work. The second section introduces several novel direct child assessment concepts emerging from the Measures Initiative. The final section describes the iterative, user-centered R&D approach the initiative is taking to develop these concepts into functional assessment products that capture the strengths and skills of all learners and can inform efforts to ensure all children have high-quality early educational experiences that foster meaningful learning. Throughout, the chapter highlights ways that research approaches and design principles of the Measures Initiative can be leveraged in assessment development in alignment with the *Principles for Assessment in the Service of Learning*, particularly in regard to assessment transparency, fairness, and design.

1 Public pre-K settings vary across states and localities. They may include public schools, child care, Head Start, and home-based child care.

A Target Product Profile as the Foundation for the Development of New Direct Child Assessment Tools

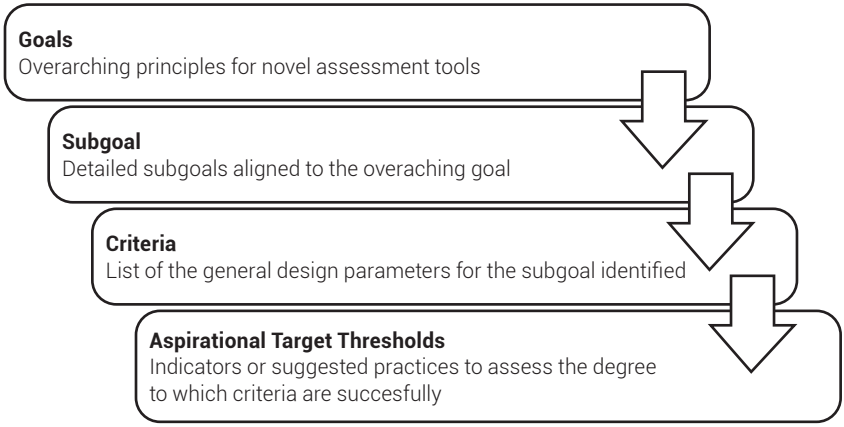
The Measures Initiative is driven by the early learning field's need for easy-to-use child assessment tools that reliably capture the widespread competencies of young learners and provide actionable insights. As an initial step, the initiative convened a wide array of users of early learning assessments to imagine the specific design features of measures meeting this vision. This process involved interviews with parents/caregivers, educators, and program leaders, as well as working sessions with academic experts in assessment, early childhood education, and developmental psychology to understand specific needs and desires for improved child assessment tools. These insights were then synthesized and organized into a target product profile (TPP). TPPs are commonly used in the health sector to outline goals, requirements, and specifications to inform the development of health care solutions such as medications and vaccines. TPPs for pharmaceutical products often include specifications for aspects like delivery mode, dosage, risks/side effects, and cost (Tebbey & Rink, 2009). Adapting this TPP approach for educational products has the potential to similarly encourage innovative solutions driven by user and market needs. Moreover, it offers a framework for reflecting on existing tools to identify their strengths, gaps, and areas for future improvement.

Developing a TPP for early learning assessments, as compared to for health products, posed unique challenges and opportunities. Whereas TPPs for pharmaceuticals typically outline parameters in set categories (e.g., dosage), the categories for key features of early learning assessments were not predefined. Similarly, whereas there are often agreed upon standards for successful medical tools (e.g., shelf stability; minimal counterindications), there were few agreed upon metrics of success for early learning assessments. The initiative's emphasis on assessment user perspectives and experiences also meant that the input received on the ideal features of child assessment tools prioritized by different engaged users was at times conflicting. For example, some preferred assessments that children could complete entirely independently to minimize teacher burden, whereas others desired tools that involve teachers to ensure they actively observe children demonstrating skills and behaviors.

Establishing a clear organizing taxonomy for early learning assessment products from thousands of user insights and comprehensive review of existing early learning assessment products and literature relied on thematic analysis. From

this approach, five areas of focus–or goals–for the next generation of early learning assessments emerged: (1) content, (2) psychometrics, (3) experience, (4) usefulness, and (5) scalability. Specific criteria elevated during interviews, focus groups, feedback sessions, and literature reviews were then organized into these five categories. Example aspirational target thresholds aligned to each criterion were generated as potential metrics to signal whether a product was progressing toward that criterion. Thresholds were designed as aspirational because, in some cases, it is unclear whether they are possible to achieve (e.g., fully offline capabilities may not be technologically feasible; it may not be possible to have brief assessments that also comprehensively cover content). Exhibit 1 describes the structure of the child assessment TPP, entitled the User-Informed Principles (MDRC & Substantial, 2022), and Exhibit 2 provides examples of subgoals, criteria, and aspirational target thresholds within each goal.

Exhibit 1. Taxonomy of the User-Informed Principles



The first goal of the User-Informed Principles describes aspirations for assessment **content** or the skills and competencies measured by early learning assessments. This section takes an expansive view on the content that child assessments should capture, emphasizing the importance of content breadth (skills across developmental domains captured) *and* depth (skills within developmental domains comprehensively reflected). Developmental domains reflect those typically included in whole-child frameworks for early learning such as the Head Start Early Learning

Outcomes Framework (Office of Head Start, 2015). This expansive perspective on assessment content stands in contrast to the common practice of focusing narrowly on assessing a subset of skills within foundational academic domains like math, language, and literacy. For example, most measures of young children's language abilities tend to center on receptive vocabulary. Although vocabulary is a foundational skill, typical vocabulary prompts testing which words children know do not reflect the full range of language skills young children are developing and ultimately need to navigate through the world. This narrow focus also advantages children who have received formal vocabulary instruction or who regularly hear and use commonly tested vocabulary words in daily life, but disadvantages those whose language strengths lie in other areas (e.g., oral storytelling). Assessments should give all children an equal opportunity to demonstrate their full range of skills and not rely on specific experiences or knowledge that are unlikely to be universal.

The remaining four goals of the TPP are relevant to assessments covering any content area. The second goal of the User-Informed Principles outlines **psychometric** properties of child assessment tools intended to ensure that they generate reliable estimates of children's skills and reflect minimal statistical bias. All assessment scores inherently include measurement error that does not reflect children's skills in the constructs intended to be measured by the tools. Therefore, this goal includes criteria for acceptable levels of measurement error in line with field standards for internal properties of educational assessments (AERA et al., 2014), but also emphasizes properties related to the fairness of the tools or their ability to generate comparable information across time, target constructs, and communities of children. A particular challenge with many existing early learning assessment tools used for both formative and summative purposes is that they rely on ratings of children's skills by an adult caregiver, typically an educator or parent. Scores from these assessments are likely to be subjective, reflecting the perspectives of those assigning the ratings, including their knowledge of early childhood development and any implicit biases about specific subpopulations of children (Cameron, McClelland et al., 2023; Gardner-Neblett et al., 2023; Russo et al., 2019). Parameters in the psychometrics goal describe features of tools that minimize the influence of rater bias on assessment scores. As a flexible framework, the TPP offers developers the opportunity to prioritize psychometric features most aligned with the intended uses of the tools they are designing.

The third goal describes parameters for the **experience** of children taking the assessment and the educators who are often responsible for using the tools to collect data. All too often, assessments are a source of stress and burden for the students and educators using them. Traditionally, direct assessments with young learners have required one-on-one educator-child sessions. For children, the repeated questioning format of these traditional one-on-one direct assessments (e.g., "What is this called?" "What color is this?") can be uncomfortable, particularly for those from communities or cultures in which these types of interactions with adults are uncommon for young children (Peña & Halle, 2011). For educators, collecting child assessment data in this way with young children can detract from their ability to engage in instructional activities in their classrooms. This is also the case when using observation-based tools, which requires extensive educator time to document and rate anecdotes on children's behaviors (Cameron, Kenny et al., 2023). This section therefore outlines parameters for direct assessment tools that are fun, engaging experiences for children to partake in and that are intuitive for educators, requiring minimal time for training and use. It also underscores the importance of these tools' integration into normal classroom routines and practices, such as free choice time and small group instruction.

The fourth goal documents properties of the data outputs from early learning assessment tools reflecting their **usefulness**. Collecting assessment data is only as meaningful as the actions the data can inform. For educators and families, assessment data can inform decisions about how to best support individual children to be successful. For early education systems, these data can inform decisions about the most effective approaches to improve early learning experiences in ways that ultimately foster positive child outcomes. Parameters in this goal underscore the importance of making data outputs timely, understandable, and actionable for a variety of purposes. It particularly emphasizes the need to make data accessible for families who typically receive limited information on their children's skills. It also highlights the potential for child assessments to serve as a conduit for collaborative communication between educators and families about children's development.

Exhibit 2.

Example subgoal, criteria, and aspirational target thresholds for each User-Informed Principles goal

Goal	Content	Psychometrics	Experience
Subgoal	1.4 Assessments capture children's skills in objective, strengths-based ways.	2.1 Assessments generate valid, psychometrically sound, and useful information for multiple purposes.	3.2 Assessments can be integrated into everyday classroom activities seamlessly.
Criteria	1.4.1 Assessments capture measures of children's learning across target domains that minimize reporter bias.	2.1.1 Assessments generate comparable construct-specific scores—with high levels of content validity as described in prior goals—across groups of 3-, 4-, and 5-year-olds.	3.2.1 Administration of the assessments can be embedded within typical pre-K routines and does not take away from other activities.
Aspirational Target Thresholds	• Assessments primarily rely on direct assessments to capture children's learning, development, and competencies. • Assessments can provide opportunities for educators to report on children's development as a supplement to direct assessment information.	• Assessments capture growth relative to a criterion (i.e., what children know and are able to do) developed specifically for priority groups with a representative sample of 3-, 4-, and 5-year-old children from diverse settings and geographic regions of the United States. • Criterion-referenced standards are available for each domain of learning and competency within-age for children ages 3, 4, and 5. • Domain scores can be compared across ages to examine growth relative to criterion-referenced standards. • Assessments yield reliable and valid scores within each age group (3, 4, 5).	• Time spent in typical instructional activities is largely unchanged (or potentially increased) before and after the take-up and implementation of the assessments in diverse pre-K settings. • Assessments are designed to be used in a variety of activities throughout the day (e.g., individual choice time or project-based time).

Exhibit 2. (continued)

Goal	Usefulness	Scalability
Subgoal	4.1 Assessments regularly generate meaningful and actionable information about children's learning, development, and competencies in separate early learning domains for multiple purposes.	5.1 Assessments are affordable for publicly funded pre-K systems and centers to administer. (Feasible price and time burden target points are currently being determined through discussions with pre-K system leaders, program administrators, and educators.)
Criteria	4.1.1 Assessments produce results that can be used to identify how children are learning and tailor instruction to support children's development.	5.1.1 There are low costs and burdens to adopt the assessments for pre-K systems and programs.

Aspirational Target Thresholds	• Assessments produce results for each child at least 6 times—or as frequently as needed by the educator to support an individual child's development—during the program year that:
	  • Can produce point-in-time holistic profiles for child development across multiple domains.
	  • Can produce reports on individual children's growth and areas for supported learning in domain-specific areas from one assessment period to the next, from the beginning of the year to the most recent assessment, and from the beginning to the end of the program year.
	  • Can produce reports on individual children's performance relative to overall classroom/group performance.
	  • Can suggest groupings of children with like abilities or mixed abilities in small groups.
	  • Can produce reports on overall classroom/group performance across multiple domains.
		• Cost of initial take-up is reasonable and feasible as agreed on by a panel of program administrators, center directors, and policymakers (costs here include the hardware and software costs to start up, and staff time to learn a new system of assessments, such as training time for educators and administrators, educators' and administrators' time to review and interpret data, and costs for IT support staff to launch, divided by the total number of children in a program or system).
		• Families, educators, and administrators in diverse early learning settings perceive the benefits of taking up and collecting the assessments to outweigh the costs of doing so after having used the assessments for at least 6 months.
		• Families, educators, and administrators are able to understand information from reports quickly and efficiently. Panel of families, educators, and administrators agree (> 80%) that implementing assessments does not detract from time spent with children or typical learning activities.

The fifth and final goal outlines aspects of tools that would ensure their **scalability** across the diverse landscape of publicly funded pre-K programs. Today most states have what is known as "a mixed delivery system" of publicly-funded programs that ranges from formal group-based settings like those within traditional public schools to informal settings like those in family child care programs typically in someone's house (Jones et al., 2020). The resources, strengths, and needs of children, educators, and administrators across these various settings greatly varies (Hanno et al., 2021). Recognizing that technology-based solutions are likely to facilitate meeting other aspects of the User-Informed Principles (e.g., covering more expansive content, improving child and educator experience), this section imagines the features of technology-enabled assessments that are accessible and implementable across contexts. It includes parameters for required infrastructure and hardware, internet connectivity, data privacy, and interoperability with other commonly used education technologies.

Together, these goals set an ambitious vision for novel child assessment products informed by the experiences and perspectives of assessment users. In contrast to some TPPs that outline minimum product specifications, the User-Informed Principles are collectively envisioned to be an aspirational target that is intentionally not prescriptive about how to accomplish individual criteria outlined within them. Assessment developers grappling with this document therefore have enormous opportunities to innovate and tackle different aspects of the User-Informed Principles with far-ranging solutions. They also are confronted by inherent tensions across and within goals: How can assessments have content that is relevant to children from different communities while also generating scores that are comparable across groups of children? How can assessments be expansive in the content they cover while also minimizing child and educator burden of collecting assessments? The User-Informed Principles are intended to empower developers to address these challenges head on with the end goal of spurring breakthrough innovations in the child assessment space.

Innovative Concepts to Advance Towards the User-Informed Principles

After developing the User-Informed Principles, the Measures Initiative moved into the next phase of work to identify and develop innovative solutions that move towards this ambitious vision for child assessments. The Measures Initiative first identified organizations with relevant expertise in early childhood, child assessment, and technology. Aligned organizations then had the opportunity to interact during in-person working sessions intended to introduce the User-Informed Principles and encourage ideation around how to further the goals outlined in the document. With these sessions as a shared foundation, organizations with complementary capacities worked collaboratively to develop initial concepts for novel direct assessments for use in public pre-K classrooms to inform instruction and decision-making. This section introduces three of the early-stage assessment concepts that emerged through this process, describing how they prioritize specific aspects of the User-Informed Principles and challenge existing assessment paradigms. All three assessments are focused initially on content areas that have traditionally been assessed through direct approaches: language, literacy, math, and executive function.

One team comprised of individuals from the Universities of Minnesota and Oregon, Aviellah Curriculum and Consulting, and FableVision Studios, an educational media and technology company, envisioned a tablet-based digital storybook-based assessment approach wherein children navigate through interactive narratives reflective of varied experiences and, while doing so, respond to prompts that capture their abilities across a broad range of skills in English and Spanish. As part of this work, the team hypothesized that individual prompts could simultaneously capture multiple skill domains. That is, the same item could shed light on children's math and receptive language at the same time. This sort of multi-dimensional assessment approach may more accurately reflect the integrated nature of children's development and skills across domains, as well as allow for measuring more content domains with minimal burden. Early-stage research with elementary schools has demonstrated the promise of engaging, multi-dimensional tools to capture children's creativity and cognitive skills (Rosen et al., 2023). This concept extends this approach by considering how it might work with younger preliterate populations, new skill domains, and within engaging storybooks.

An alternative hypothesis proposed by another Measures Initiative team is that short tablet-based games can expand content measured while also improving children's experiences taking assessments. This hypothesis is posed by Khan Academy Kids, an education technology non-profit that currently has a free, widely used learning application containing thousands of interactive learning activities for young children ages 2 to 8 based in the Head Start Early Learning Outcomes Framework and Common Core Standards. Khan Academy Kids' proposed approach of creating short engaging assessment tests styled after their existing learning activities seeks to break down the silos between play, learning, and assessment. As part of their work, this group is also considering whether their short, embedded assessment tasks can yield both formative and summative insights. That is, can these brief assessment modules be flexibly administered, adapted, and scored to both inform teacher practices and instruction, as well as monitor children's learning and generate large-scale summative assessment scores to support continuous program improvement?

A third assessment concept explored through the Measures Initiative is whether children's skills can be accurately captured as they navigate through physical books and respond to items about book content. Picture books are commonplace in early childhood classrooms and are used throughout the day during whole group circle time, small groups, and independent learning centers. This team, led by learning technology and educational experts at Kibeam and Mighty Play, has developed a book-based assessment learning and reporting system. Children navigate through books using a small handheld device or "wand" developed by Kibeam. The wand is equipped with sensors that read images and text on a page and a speaker that can read aloud text and interactive assessment prompts on each page. This physical technology not only allows preliterate readers to independently engage with written text in new ways, including kinetic interaction, it also offers a potential approach for asking and recording children's responses to prompts related to content on pages. The team proposed exploring the feasibility of continuously collecting data to enable embedded, ongoing assessment through diverse picture books. This novel tool represents a potentially joyful, engaging way of collecting data naturalistically as children engage with activities (i.e., book reading) that they would already typically do in pre-K classrooms.

Each of these assessment concepts represents a unique hypothesis about how to improve data collection on young children's skills and abilities. They advance different assessment methods (i.e., digital storybooks, technology-enabled short learning activities, and a handheld book reading device) intended to engage children and reduce teacher burden when conducting assessments, while also developing content for emergent bilingual preschoolers learning Spanish and English. Across the initiative, each developer team has also considered whether recent technological developments might buoy their efforts to progress toward the User-Informed Principles. Rapid advancements in artificial intelligence (AI), in particular, offer unique opportunities to address various limitations of existing early learning assessments. These technologies could help efforts to broaden and deepen assessment content. For example, generative AI models could quickly develop expansive assessment content or vignettes within which to embed assessment items. Improvements in automated speech recognition (ASR) capabilities may mean that young children's expressive language can be accurately captured and analyzed, further expanding the types of content direct assessments are able to measure. These new technologies may also have implications for educators using the tools, providing them with in-the-moment guidance on how to more efficiently and objectively capture, analyze, and react to data.

Despite the promise of AI for improving various aspects of early learning assessments, these new capabilities remain largely untested and could unintentionally exacerbate existing assessment challenges without careful consideration and monitoring (Ho, 2024). AI models rely on training datasets to learn how to process and generate information. The perspectives and experiences reflected in the training datasets will therefore filter into what is produced by AI models. This could create challenges for the quality of insights generated from ASR-based assessment prompts if ASR models are generated with a limited set of voices. For example, models built with training data comprised of adult speech samples are unlikely to reliably capture the unique speech patterns of young children (Patel & Scharenborg, 2024). Relatedly, models built with datasets that prioritize specific dialects may not accurately capture the speech of those who speak excluded dialects (Wassink et al., 2022), likely resulting in incorrect estimates of certain communities' expressive language abilities. These risks underscore the importance of examining the consequences of integrating AI technologies into assessment products through research conducted in partnership with diverse communities.

Active Engagement with Assessment Users in Iterative R&D

The Measures Initiative employs an iterative research and development approach to ensure early-stage tools and technologies building from these assessment concepts are progressing in line with the priorities and needs of assessment users. Turning an assessment concept into a functional product requires frequent, ongoing decision-making on the part of assessment developers. These countless decisions range from the macro (e.g., the content domain(s) to assess) to the micro (e.g., the color schemes to use in data dashboards; the words to use to direct children to the next assessment item). In traditional assessment development paradigms, assessment developers largely make these decisions based on their personal experiences and expertise. In the absence of external voices, the developers' perspectives are then naturally reflected in the assessment product and the data that come from it. For example, as noted above, measures of young children's language skills often focus on receptive vocabulary tasks that examine children's ability to understand a list of words spoken out loud, meaning that these tools reflect a narrow conceptualization of language skills (excluding expressive, syntactical, and social components; Portilla & Iruka, 2024). Despite the narrow perspectives drawn on during initial development and validation work, these measures are now commonly used across the United States in socio-demographically diverse samples for research and monitoring purposes. Given how the tools were developed, differences in scores across subgroups on these vocabulary measures may indicate differences in the relevance of the tools across populations rather than underlying skill differences. That is, these tools afford some but not all children the opportunity to show what they are learning and doing in their homes and communities, undermining the fairness of assessments.

In contrast to this dominant approach, the Measures Initiative has sought to integrate user perspectives by drawing on the expertise of communities served by today's public pre-K programs at every step of the tool development process. Over time, tools developed through the initiative have evolved from concept ideas to prototypes to functional assessment products with comprehensive item banks that cover multiple content domains. This means that the assessment components and functionalities that the developers are co-designing and pressure testing with assessment users over time exponentially grow. The quantity and variation of assessment features being developed also means that no one research activity is likely to be sufficient to build evidence on the extent to which they meet user

needs, accomplish criteria outlined in the User-Informed Principles, or align with field standards for educational assessments. That is, the same research activity is unlikely to yield meaningful insights on a set of 100 math items as on a new data dashboard prototype.

With these considerations in mind, the Measures Initiative has implemented an iterative, cyclical, and phased R&D approach involving assessment users representing a range of communities in a variety of research activities intended to generate insights and evidence on different aspects of early-stage assessment tools. Assessment user groups represented across different research activities include pre-K students, educators, parents/caregivers, and program administrators. These proximal assessment users–those expected to take assessments or collect and use data from the tools–are recruited to Measures Initiative activities through the initiative's close partnerships with local agencies that have longstanding relationships with early education programs (e.g., recruitment and referral organizations; technical assistance providers).

By working with organizations with system-wide purview, the initiative engages individuals from a broad range of program types across different geographic contexts. The pre-K landscape is notoriously fragmented in the United States with families and children relying on a variety of publicly funded and subsidized education and care types (e.g., Head Start, public school-based pre-K, community-based centers, and family child care programs). Yet, contemporary early childhood research rarely reflects this diverse, patchwork landscape, often constrained to a narrow set of classroom-based programs predominantly in cities (Jones et al., 2020). In the Measures Initiative, having a broader perspective on the early childhood landscape is particularly important for exploring how the tools might work across early learning programs with educators who have different professional experiences (e.g., education levels, certification), instructional supports (e.g., coaching, curricular materials), and technological resources (e.g., high speed internet).

The initiative has also built a network of field leaders who bring practice, policy, and academic expertise to engage in its R&D process. Although these individuals may be less likely to use the tools directly than those recruited from programs, they are often consumers of assessment data from these tools and make or influence decisions about the assessment tools used in publicly funded pre-K programs. Moreover, they each bring valuable expertise relevant to different

aspects of tool development. For example, the academic experts typically bring extensive knowledge on children's skill progressions in early childhood, with many contributing to state and federal early learning standards that guide what early learning programs are hoping their students will learn. In contrast, many of the practice and policy experts have experience acquiring and rolling out new assessments at the systems level.

Research activities with this broad range of assessment users are then meant to provide opportunities for authentic co-design between developer and user, as well as yield insights about tools' progress towards the User-Informed Principles' goals (i.e., content, psychometrics, experience, usefulness, and scalability). Importantly, not all users engage in all activities or are asked to weigh in on all aspects of the tool. For example, pre-K students are important partners for designing the experience of using the tool in real world classroom conditions but are appropriately not tapped to evaluate an assessment's scoring procedures. Similarly, not all research activities occur at every phase of tool development or are expected to provide insights on every aspect of the tools. As tools become more developed with increased functionalities, additional research activities are layered on to more comprehensively evaluate tool features. Below is a list of the types of research activities that the Measures Initiative has integrated into its R&D process, describing the user groups engaged and the assessment areas interrogated through each activity:

Focus groups and feedback sessions: Starting with their concept designs, developers have met with small groups of users to share materials they are developing and iterate on those materials with users. Most often these groups are comprised of educators and administrators or parents/caregivers, but teams also occasionally meet with practice, policy, and content experts either in small groups or individually. These conversations often focus on getting users' thoughts on assessment content, assessment interface or plans for how the tools might be used in classrooms (experience), or data dashboards (usefulness).

Content vetting: Once developers have initial assessment content (e.g., construct maps, item banks), teams of academic experts with deep knowledge of early childhood education, child development, and measurement holistically evaluate the assessments, providing thoughts on each tool's general approach to content, experience, and usefulness. They also review each individual assessment item,

noting whether they believe it is capturing the intended domain/subdomain and is developmentally appropriate for pre-K students.

Cognitive interviews: Once developers have early-stage prototypes of their assessments, they can conduct one-on-one cognitive interviews with children and educators to observe and learn how they navigate the tools (experience) and interpret the item prompts (content). In the case of educators, these interviews also often include having educators review data outputs to evaluate whether the outputs allow educators to accurately interpret scores and make well-supported instructional decisions from the data (usefulness).

User testing: Once developers have functional early-stage assessment tools, they can have small samples of educators use their tools in real world pre-K classrooms over an extended period of time (e.g., several weeks or months). This gives developers the chance to iterate on the training, implementation materials, and ongoing supports they provide to educators using their tools. It also provides insights into what it is like for children and educators using the tools (experience) and whether these experiences differ across settings (scalability).

Psychometric analysis: Assessment data collected through user testing is used to evaluate the quality of the information and, ultimately, scores coming from the tools. At the earliest stage when user testing is constrained to small samples comprised of a handful of classrooms, item-level data are examined to ensure items have varying levels of difficulty and are not likely to produce scores that suffer from ceiling or floor effects. Over time, as user testing involves larger samples, psychometric analyses can become more complex, examining psychometric properties like scale reliability, differential item functioning, and correlations with established measures. They can also be used to build evidence on scoring procedures like stopping rules or computer adaptive testing approaches that can help reduce the number of items children must complete.

Over time, developer teams compile insights from these various co-design activities repeated with many assessment users. In some cases, suggestions and solutions raised in these activities are quickly implementable. For example, when children tested out several of the tools during cognitive interviews, they often forgot to click the on-screen or physical button required to advance to the next item. Based on that observation, developers were able to quickly program consistent

prompts reminding children how to move forward in the assessment and then were able to determine whether those prompts helped keep children advancing through items while observing them in their next round of interviews. Other times, teams are required to synthesize and digest multiple and at times contradictory inputs on the same assessment features. This sometimes occurs within research activities: focus group participants disagree with each other or content vetters rate the same item differently. Other times, teams get contrasting feedback from different research activities. User testing might illuminate the need for assessments to be shorter to more seamlessly integrate them into classroom schedules and sustain child engagement, while psychometric analyses might suggest additional items are needed to yield more reliable estimates of children's skills. Teams are encouraged to directly grapple with these contradictions, recognizing that although there is unlikely to be a single correct path forward, this iterative co-design and development process offers opportunities to rigorously test innovative solutions and understand how assessment users experience and perceive them. Teams are also encouraged to critically reflect on how the solutions they initially propose might reflect their own experiences, expertise, and preferences rather than those of the assessment users they have partnered with. This is in service of continually seeking to prioritize the experiences of those who will ultimately use and be affected by the tools such that they can be usable, useful, and generate accurate insights on all children's abilities.

So far, this approach has brought particularly valuable insights into assessment design features that support young learners' ability to successfully demonstrate what they know and can do. Traditional direct assessment approaches for older students—like pen and paper or computer-based assessments—are not developmentally appropriate for pre-K-aged children who are often preliterate and lack computer skills to navigate through a computer-based assessment (e.g., mouse handling, typing on a keyboard). Consequently, the data collection approaches (i.e., tablet-based or handheld device) tested through the Measures Initiative are relatively new.

Even with strong grounding in developmental science, it can be challenging to predict all the ways young children might interact with new technology-based direct assessment interfaces that might affect data quality. For example, we did not anticipate that young children, when given the chance, would complete the same assessment game or story multiple times without explicit controls preventing

them from doing so or that some children would figure out how to complete assessments under other children's profiles. By observing these behaviors through user testing, developers are able to design new approaches to ensure children can focus on the assessment task at hand rather than being distracted by unintended ways to engage with the tool.

Similarly, developers have begun to identify assessment features that foster children's motivation and continued engagement during administration. Children appreciate the opportunity to have choice, such as being able to select which assessment tasks to complete first or the character that provides instructions during the assessment. Varied, child-friendly design elements—including high-contrast colors and familiar assets—appear to encourage sustained focus during tasks. During user testing, children remarked about familiar or favorite items used in assessment prompts (like manipulatives in patterning tasks). Varying item response modalities, such as those that allow children to verbally or kinesthetically respond to prompts, can also keep children engaged and prevent mindless tapping through items. Although, importantly, these novel response approaches must come with clear and simple instructions about what to do. Children not only need explicit guidance on how to respond to assessment prompts, but also how to navigate the assessment application interface. For example, digital assessments often include a button to click to advance to the next item, which although intuitive for older children and adults, is not familiar for most young children. All assessments coming out of the Measures Initiative include brief training modules to acclimate children to the maneuvers they will need to know to navigate through the assessments.

User-testing has also provided insights into how to make assessments more usable for early educators. Given that pre-K classrooms are dynamic environments with lots going on, educators have requested the option to pause assessments midway to allow children the opportunity to start back where they left off rather than having to repeat completed items. This could accommodate common short interruptions like bathroom breaks. Future testing through the Measures Initiative will explore whether these types of short pauses affect student performance on assessments. Teachers also requested that initial training materials include more concrete guidance on how to set up and use the assessments in their classrooms. This includes how to store, charge, and turn on technology; how to connect devices to the internet; and how to use the tool during different instructional formats (e.g., small groups, centers). These early insights illuminate the importance of

considering user perspectives in assessment design to align features and supports with what will work in real world settings and give children the best opportunity to demonstrate what they know and can do.

Conclusions: Towards a New Paradigm for Assessment Development

Young children have an incredible capacity to build skills and learn new things. The ability to foster this development in early learning programs rests on understanding where children are in their development to best tailor supports and how they grow over time to unearth the best ways to help them advance. This chapter described the ambitious goals of the Measures for Early Success Initiative to build better tools for early education programs that can support high quality early learning experiences for all young children. Beyond the goal of this work to produce new assessment tools, the initiative also advances a vision for assessment development that centers the experiences and perspectives of assessment users rather than assessment developers. This co-design and co-development process can result in tools that are appropriate for use in a broader range of communities and settings. It can also encourage greater transparency about assessment tools by generating clear evidence on tools' strengths and limitations from R&D activities with assessment users. No one assessment tool is likely to meet every user's needs or every aspect of the User-Informed Principles introduced in this chapter, but being clear about what tools can and cannot do for different users can help ensure tools are not used in the wrong ways. All children deserve the opportunity to be able to show what they know and can do and have those gifts recognized by the adults in their lives. Assessments that capture, recognize, and connect to resources that foster these gifts are important starting points.

References

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin, 145*(12), 1103–1127. https://doi.org/10.1037/bul0000212

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Cameron, C. E., Kenny, S., & Chen, Q. H. (2023). How Head Start professionals use and perceive Teaching Strategies Gold: Associations with individual characteristics including assessment conceptions. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2022.103931

Cameron, C. E., McClelland, M. M., Kwan, T., & Starke, K. (2023). HTKS-Kids: A tablet-based measure of self-regulation to equitably assess preschoolers' school readiness. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1202239

deMonsabert, J., Brookes, S., Coffey, M. M., & Thornburg, K. (2021). Data use for continuous instructional improvement in early childhood education settings. *Early Childhood Education Journal, 50*(3), 493–502. https://doi.org/10.1007/s10643-021-01168-3

Gardner-Neblett, N., De Marco, A., & Ebright, B. D. (2023). Do Katie and Connor tell better stories than Aaliyah and Jamaal? Teachers' perceptions of children's oral narratives as a function of race and narrative quality. *Early Childhood Research Quarterly, 62*, 115–128. https://doi.org/10.1016/j.ecresq.2022.07.014

Hanno, E. C., Gonzalez, K. E., Jones, S. M., & Lesaux, N. K. (2021). Linking features of structural and process quality across the landscape of early education and care. *AERA Open, 7*. https://doi.org/10.1177/23328584211044519

Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics*, 10769986241248771. https://doi.org/10.3102/10769986241248771

Hsueh, J. (2021). *Challenge and opportunity: Equitable pre-k measures for early learning*. https://www.mdrc.org/work/publications/challenge-and-opportunity

Im, H. (2017). Kindergarten standardized testing and reading achievement in the U.S.: Evidence from the early childhood longitudinal study. *Studies in Educational Evaluation*, *55*, 9–18. https://doi.org/10.1016/j.stueduc.2017.05.001

Jones, S. M., Lesaux, N. K., Gonzalez, K. E., Hanno, E. C., & Guzman, R. (2020). Exploring the role of quality in a population study of early education and care. *Early Childhood Research Quarterly*, *53*, 551–570. https://doi.org/10.1016/j.ecresq.2020.06.005

McCormick, M., & Mattera, S. (2022). *Learning more by measuring more: Building better evidence on pre-k programs by assessing the full range of children's skills*. https://www.mdrc.org/work/publications/learning-more-measuring-more

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., Yang, R., Koepp, A., & Shonkoff, J. P. (2017). Impacts of Early Childhood Education on Medium- and Long-Term Educational Outcomes. *Educational Researcher*, *46*(8), 474–487. https://doi.org/10.3102/0013189X17737739

MDRC & Substantial. (2022). *User-Informed Principles: Developing Assessments for All Early Learners*. https://www.mdrc.org/work/publications/user-informed-principles

Office of Head Start (2015). *Head Start Early Learning Outcomes Framework*. Administration for Children and Families, U.S. Department of Health and Human Services. https://headstart.gov/interactive-head-start-early-learning-outcomes-framework-ages-birth-five

Patel, T., & Scharenborg, O. (2024). Improving End-to-End Models for Children's Speech Recognition. *Applied Sciences*, *14*(6), 2353. https://doi.org/10.3390/app14062353

Peña, E. D., & Halle, T. G. (2011). Assessing preschool dual language learners: Traveling a multiforked road. *Child Development Perspectives*, *5*(1), 28–32. https://doi.org/10.1111/j.1750-8606.2010.00143.x

Portilla, X. A., & Iruka, I. U. (2024). *Advancing equity in pre-k assessments: Elevating the strengths of children from racially and linguistically marginalized backgrounds*. https://www.mdrc.org/work/publications/advancing-equity-pre-k-assessments

Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multi-dimensional approach for enhancing and measuring creative thinking and cognitive skills. *The International Journal of Information and Learning Technology, 40*(4), 334–352. https://doi.org/10.1108/IJILT-12-2022-0227

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, *48*, 14–25. https://doi.org/10.1016/j.ecresq.2019.02.003

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games.* The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning. MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001

Tebbey, P. W., & Rink, C. (2009). Target product profile: A renaissance for its definition and use. *Journal of Medical Marketing*, *9*(4), 301–307. https://doi.org/10.1057/jmm.2009.34

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Society for Research in Child Development. https://www.fcd-us.org/wp-content/uploads/2016/04/Evidence-Base-on-Preschool-Education-FINAL.pdf

Open Badges as Assessment Innovation: From Digital Media Revolution to AI-Enabled Futures

Constance Yowell and Girlie C. Delacruz

Introduction—Movement Grounded in Experiment

It's 2025. We have been graciously invited to contribute to this extraordinary volume a brief introduction to the topic of Open Badges as educational assessment. To ground this essay, we begin twenty years ago in 2005, in part because, today, as the AI revolution takes off, envelopes us, and demands our attention, we are regularly reminded of the middle and late 00s (or aughts)—another time when a somewhat similar revolution—in digital and social media—took off. Open Badges, and the story of their origin and evolution, may provide a useful window for considering the current opportunities and challenges for assessment innovation.

Traditional forms of assessments rarely capture the richness of real-world competencies and Open Badges were designed to fill that gap. The concept of badges as recognizing discrete, stackable demonstrations of skill is not new. As Baker and Delacruz (2015) note the Boy Scout merit badge system, established in 1911, pioneered breaking down complex achievements into specific, demonstrable skills through authentic tasks. This framework laid the groundwork for today's focus on competency-based learning (Patrick & Sturgis, 2013), where students advance based on demonstrated mastery rather than seat time or test scores. The discrete, stackable nature of merit badges mirrors the structured attainment levels found in qualification frameworks across the United Kingdom, Australia, and New Zealand. A century later, digital badge systems are using technology to recognize real-world skills on a much larger scale.

Like the "Napster moment" when the file-sharing service disrupted the music industry by demonstrating new possibilities without immediately replacing existing systems, Open Badges have pointed toward transformative possibilities while grappling with deeper structural challenges in credentialing and recognition systems. The intentional design of Open Badges with their roots in rigorous educational theory, robust metadata, and a commitment to equity sets the stage for their practical application across diverse educational and workforce contexts.

Connecting Open Badges to Principles of Assessment Innovation

In 2005, the John D. and Catherine T. MacArthur Foundation launched an initiative in *Digital Media and Learning*, eventually investing $250 million over a decade to support research and the design of new approaches to learning. We write as co-architects of the Open Badges project infrastructure—one of us a program officer at the MacArthur Foundation, the other a grantee involved in the implementation effort at the field level. Our goal is not to defend the work, but to reflect on the design intentions and future value of Open Badges. This period revealed to us and others involved the emerging potential of digital media and the Internet to transform learning from its traditional focus on content consumption—what James Gee evocatively refers to as "a fetish on consumption"—to more participatory and production-oriented forms (Gee, 2003). It also became clear that traditional forms of recording and signaling learning—primarily content mastery attested by grades or diplomas—did not capture much of what mattered to learners, nor did they reflect the realities of digital participation.

By 2010, the marriage of deeper learning principles with the technical architectures of the Internet was not just possible, but necessary. In 2011, the Mozilla Foundation, Peer 2 Peer University, and the MacArthur Foundation, released the foundational Open Badges white paper outlining the three core components of a badge infrastructure: the badges, underlying assessment practices, and technological standard and metadata framework that enable cross-contextual use (Mozilla Foundation & Peer 2 Peer University, 2011).

From the start, Open Badges were intentionally crafted to align with cutting-edge research on pedagogy and assessment. The early design teams collaborated closely with leaders in game-based learning and equity-driven assessment—many of whom have contributed to this Handbook series—to ensure the metadata and badge infrastructure reflected the following overarching goals. Open Badges use an

argument-based approach (Kane, 1992) to establish validity, triangulating evidence and analysis to support validity claims within specific contexts. A badge's credibility depends on the quality and transparency of the evidence behind it. By grounding the design of Open Badges in established frameworks like Evidence-Centered Design (Mislevy, Almond, & Lukas, 2003) and Model-Based Performance Assessment (Baker, 1997), then encoding these principles into machine-readable metadata, Open Badges made technology essential to establishing validity in badge-based assessment.

To that end, the Open Badge Standard was designed to include high quality and transparent evidence of learning and performance. The metadata specification—the "bones" of a badge—was created to include, among other things:

- **Achievement descriptions** that detail what the badge represents, its context and specific achievements;
- **Criteria and requirements** that detail what must be met and completed to earn the badge;
- **Evidence** that provided examples of the work or documentation justifying the award of the badge;
- **Standards Alignment** that included a reference to educational or industry frameworks.

This attention to transparency, transferability, motivation, structure, adaptation, equity, and quality echoes the seven animating principles of this volume:

—***Principle 1***: With an emphasis on transparency, every badge includes clear descriptions, explicit criteria, and links to evidence—making assessments understandable to all stakeholders.

—***Principle 2***: With an emphasis on transfer and explicit focus, badges aimed to document skills and outcomes in ways that could be meaningful across diverse settings.

—***Principle 3***: With an emphasis on motivation and engagement, the flexible design was intended to ensure they were "owned" by the learner and supported reflection through self-curated learning pathways.

—***Principle 4***: With an emphasis on modeling expectations, Open Badges can scaffold and represent structured learning progressions—horizontal or vertical—across time.

—***Principle 5***: With an emphasis on feedback and adaptation, Open Badges could incorporate immediate feedback through iterative tasks and adapt to various forms of learning and assessments.

—***Principle 6***: With a driving emphasis on equity, Open Badges enable credentialing of skills gained in community, informal, or workplace settings—not just traditional academic venues—broadening participation and valuing often marginalized forms of learning.

—***Principle 7***: Emphasizing quality and validity of evidence, each Open Badge embeds access to evidence, issuer reputation, and standard alignment.

With this theoretical and technical foundation in place, we turn now to real-world implementations that test these principles in practice.

Use Cases—Learning from Experience

In today's evolving workforce, valid credentials serve as powerful levers to unlock opportunity. We provide two examples to exemplify this potential: the "This Way Ahead" Gap Inc. workforce preparation program and a badge-to-credit initiative in partnership with Southern New Hampshire University (SNHU).

Each of these examples was a project run by LRNG, a nonprofit, supported by the MacArthur Foundation and established by the authors in 2015 to design and implement "badged" pathways of learning for youth across cities and their communities. The LRNG badge and pathway platform reframed learning as a connected ecosystem, partnering with schools, city agencies, businesses, community organizations, libraries, and museums. Two core elements of the LRNG platform were playlists, which were narrative collections of one or more online or in-person experiences (XPs) stitched together into a compelling media-rich narrative around a common theme. Learners could also earn an LRNG badge to provide verifiable evidence of a substantive learning outcome of an organization's choosing. Badge credibility rested on community norms and shared values. Sometimes badge issuers restrict acceptable evidence types, based on what was appropriate for the learning experience and what counts within

that community. Other times, learners had full discretion over what to submit that counted as evidence. As such, the creation and empirical inspection of the validity argument put primary emphasis on front-end specificity in collaboration with relevant stakeholders including students, employers, and curriculum designers.

LRNG badges also integrated community membership and uptake as part of the validity argument. Badge metadata recorded the issuing organization, making it clear whose norms and values underlie the credential. Ecosystem members could share and re-issue badges, creating networks of endorsement, bolstering their credibility. When multiple organizations recognized and even re-issued the same badge, they collectively affirmed the value of both the credential and its supporting evidence.

The LRNG Platform made the learning network visible, surfacing who else had adopted each badge and reinforcing each badge's validity through community demand. This convergence of structured metadata, evidence artifacts, community endorsements, and transparent inspection demonstrated how technology could weave evidence and inference into a single, interoperable credential.

Each of these examples illustrates how the foundational principles and architecture of Open Badges have been translated into practice, and how badges, grounded in rigorous assessment design can reliably signal learner competencies and open pathways to employment and higher education.

"This Way Ahead" Digital Pilot

The This Way Ahead Digital Pilot (TWADP) brought together Gap Inc., community-based partners, and LRNG to create a suite of Open Badges that qualified young people to interview at Gap retail stores. Drawing on Gap Inc.'s *This Way Ahead* curriculum and insights from interviews with human resource specialists, store managers, and regional managers, LRNG focused on teaching and assessing three core competencies for entry-level sales associates: Teamwork, Conflict Resolution, and Punctuality.

We linked each badge to behavioral objectives, tasks, evidence, and rubrics in a model-based framework. For each of these learning outcomes, we specified the tasks learners would complete, the evidence they needed to submit, and the rubric criteria for scoring. Gap Inc. staff reviewed the framework to confirm that it accurately reflected the targeted competencies and that the artifacts learners submitted constituted valid, appropriate evidence of mastery for each of the learning outcomes.

One illustrative activity asked learners to recount a personal example of teamwork or conflict resolution. Gap Inc. staff reported that strong candidates could effectively articulate how they have used these skills in their lives. We asked learners to draft a concise 2–3 sentence written response and then record a short video practicing their delivery. This two-step task guided learners to draw on examples of using these skills in diverse contexts such as at school, with friends or family, on sports teams, or in clubs. It then had them practice voicing their responses aloud, mirroring how they would share those examples in a real interview. Learners reported feeling more confident in interview settings, and many badge-earners subsequently received job offers from Gap Inc.

This pilot demonstrated how the LRNG platform's Open Badges integration codified the assessment argument directly into each badge. Written reflections, video recordings of learners practicing their responses, and answers to scenario-based quizzes were logged. This data formed the raw material for each badge's evidence field, ensuring that every submission was timestamped, verifiable, and tied directly to the competency being assessed. Once the learner's scores and human ratings met the badge-award thresholds, a rule engine triggered the badge assertion and a badge was awarded which contained the scored artifacts, as well as the seal of authority which denoted Gap Inc. as the issuer, making the entire evidentiary chain visible in the LRNG dashboard. Learners and badge consumers (e.g., future hiring managers, nonprofit partners) could inspect how each claim was supported, making the LRNG badge a self-contained, interoperable argument of competency.

Badges-to-Credit Initiative

LRNG, One Summer Chicago (City of Chicago's summer youth employment program), and Southern New Hampshire University (SNHU) collaborated to demonstrate how informal learning can be translated into formal college credits. Together, they identified a set of playlists and badges that could be awarded credit equivalency through the process of prior learning assessment. Prior learning assessment comprises the processes and practices of determining if knowledge, skills, and abilities gained in a variety of settings may warrant consideration of college credit. For this work, SNHU used the Global Learning Qualifications Framework (SUNY Empire State College, 2014) to determine course equivalency, evaluating playlists developed by the youth serving organization, scoring rubrics, and samples of student submissions. As a result, 36 playlists and badges were

identified to count toward 19 course credit equivalencies. This canonical set comprises career readiness, design, and coding playlists and badges. For each of the identified SNHU Competencies or Courses that map onto a set of LRNG badges, we created an SNHU meta-badge on the LRNG Platform, to be automatically issued when an LRNG learner earns all the associated LRNG badges.

Because the LRNG Badges were developed using a model-based framework each badge embedded an explicit chain of reasoning among the learning outcomes, required evidence, and scoring criteria direction into its metadata. This self-contained assessment argument enabled the SNHU team to transparently inspect every badge's linked artifact, rubric scores, and badge issuer to ensure they could verify competency before awarding course credit.

The value of this work is that it fundamentally breaks the singular control of schools in defining learning that counts. Young people were able to participate in robust experiences in summer youth employment, after-school programs, entrepreneurship experiences and more that occur anywhere, anytime while simultaneously building their work and college portfolios.

Badges provide pathways to opportunity that can bypass the lengthy timelines required for degrees or certifications, allowing learners to demonstrate competency and gain recognition as soon as skills are mastered. Such flexibility can enable us the opportunity to redesign and reimagine pathways to social mobility that are grounded in the needs and interests of each young person. It also brings the possibility of college and a meaningful career closer to our young people, enabling them to see that their learning experiences build a clear and immediate path toward higher education.

What We've Learned, What Remains Unfinished

Fifteen years since their launch, with inspiring examples such as those shared here and many others, it is possible to feel extremely hopeful and optimistic about the potential for Open Badges to enable the equitable scale of high-quality learning and innovative assessments. It is also possible to experience ambivalence, and wonder if rather than enabling transformation at scale, their influence has more closely resembled that “Napster moment” as a disruptive innovation that unsettled established norms, provoked new conversations, and pointed toward what might be possible, without resolving deeper structural challenges.

The Open Badge standard (IMS Global Learning Consortium, 2015) and its associated infrastructure clearly create the necessary digital foundations for innovative and equitable assessment, as articulated by Gordon and Rajagopalan (2016) and the volume's authors. In contrast with grades, transcripts, and resumes, which reinforce traditional conceptions of achievement, badges offer the architecture for recognizing diverse and meaningful learning. Over recent years, Open Badges have undergone significant technical upgrades with version 3.0's enhanced security features that make each badge cryptographically verifiable creating tamper-proof digital credentials. At the same time, the Comprehensive Learner Record standard evolved to version 2.0 that can collect and organize multiple credentials into a single, authenticated record that learners own and control. Together, these developments align with the establishment of a global standard for Learning and Employment Records, which integrate verifiable micro-credentials into interoperable learner-controlled portfolios, advancing both portability and trust across educational and workforce ecosystems (1EdTech Consortium, 2024; 1EdTech Consortium, 2025; Institute of Electrical and Electronics Engineers, 2024).

Yet, the infrastructure alone has been insufficient to drive systemic change: while Open Badges can encapsulate granular evidence of learning, their widespread use is marked by fragmentation and inconsistency. Features like robust metadata, the organization of badges into coherent, stackable pathways, and systematic unlocks of new opportunity for learners remain only partially realized. They haven't become the engine for assessment innovation we once hoped for—at least, not yet.

There has, nonetheless, been significant cultural impact. Startups centered on digital credentials, portfolios, and learner wallets underscore a shift in narrative about the future of learning pathways. Millions of badges have been issued worldwide (1EdTech Consortium & Credential Engine, 2023). Universities regularly produce micro-credentials as part of their curriculum, and the language of modular, "stackable" credentials is commonplace in higher education circles (Coursera, 2024). For example, Western Governors University uses a unified credential framework and extensive rich skills descriptor library to integrate digital badges with degree pathways, allowing students to demonstrate competencies incrementally rather than waiting for program completion (Western Governors University, n.d.-a; Western Governors University, n.d.-b). This capacity for badges to demonstrate competency and gain recognition as soon as skills are mastered aligns with the rise of a skills-based economy. Recent research indicates that 81%

of employers believe skills should be prioritized over degrees, and 95% of university leaders expect micro-credentials to become a standard feature within most degree programs (HolonIQ, 2023).

These developments speak to a growing awareness of—and demand for—alternative recognitions of learning, even if they have not (yet) led to truly systemic assessment innovation aligned with the seven animating principles.

The (Still) Missing Links—and Why AI Might Matter

If Open Badges are ever going to matter for assessment, they'll have to serve as bridges. The infrastructure was intended to connect learning experience, skill, evidence, innovative assessment, and, finally, opportunity. In theory, all of that can be rendered explicit in the badge metadata, aligned to the seven design principles described earlier.

Open Badges serve a de-coupling function: they enable curriculum to be chunked into smaller, more isolated pieces. It is less common to find Open Badges actively linking coherent, living pathways, as demonstrated in the Badges-to-Credit example, where they linked a learning pathway across institutions (youth development organization to college credit and to job opportunities).

This is where AI enters the story, offering the most credible chance in years to close the gaps. Large Language Model powered systems make performance-based assessment scalable by delivering real-time, personalized feedback, analyzing student work processes, and adapting tasks on the fly, functions that were once costly and labor-intensive to implement at scale. With Open Badge Standard 3.0 (1EdTech Consortium, 2025) able to ingest diverse technology files, these assessments can be formally captured and verified as digital credentials. Automating data capture, matching evidence to criteria, mapping pathways, and even identifying opportunity—all of these become possible with the right human-centered application of AI. Real integration could at last take shape, relieving users of the burden and letting badges begin to function as intended. But it's an open question whether that future will materialize, or if badges will remain a prototype for what comes next.

References

1EdTech Consortium & Credential Engine. (2023). *Open Badge Count 2022: Findings*. 1EdTech Consortium.

1EdTech. (2024). Six steps to a skills-based ecosystem: A playbook for action. https://www.1edtech.org/resource/credentials-case-studies

1EdTech Consortium. (2024). Open Badges 3.0 specification. https://www.imsglobal.org/spec/ob/v3p0/

1EdTech Consortium. (2025). Comprehensive Learner Record standard 2.0. https://www.imsglobal.org/spec/clr/v2p0

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, *36*(4), 247–254.

Baker, E. L., & Delacruz, G. C. (2015). Badges and skill certification. In J. M. Spector (Ed.), *Encyclopedia of Educational Technology* (pp. 71–74). Sage Publications.

Coursera. (2024). *Micro-Credentials Impact Report 2024*. Coursera.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy.* Palgrave Macmillan.

Gordon, E. W., & Rajagopalan, K. (2016). *The testing and learning revolution: The future of assessment in education*. Palgrave Macmillan.

HolonIQ. (2023, September 28). *The future of post-secondary education in the US*. https://www.holoniq.com/notes/the-future-of-post-secondary-education-in-the-us

IMS Global Learning Consortium. (2015). Open Badges Specification v1.1. https://www.imsglobal.org/sites/default/files/Badges/OBv2p0/history/1.1-specification.html

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033–2909.112.3.527

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Practice, 1*(1), 3–62.

Mozilla Foundation, Peer 2 Peer University, & MacArthur Foundation. (2011). *Open Badges for Lifelong Learning: Exploring an open badge ecosystem to support skill development and lifelong learning for real results such as jobs and advancement* [White paper]. Mozilla Foundation. https://wiki.mozilla.org/images/5/59/OpenBadges-Working-Paper_012312.pdf

Patrick, S., & Sturgis, C. (2013). *Quality and equity: The promise of competency-based education*. International Association for K–12 Online Learning (iNACOL). https://aurora-institute.org/resource/quality-and-equity-the-promise-of-competency-based-education/

SUNY Empire State College. (2014). *Global Learning Qualifications Framework*. https://sunyempire.edu/global-learning-qualifications-framework/

Western Governors University. (n.d.-a). *The WGU Skills Library*. https://www.wgu.edu/lp/general/wgu/skills-library.html

Western Governors University. (n.d.-b). *Unified Credential Framework*. https://www.wgu.edu/content/dam/wgu-65-assets/western-governors/documents/skills/WGU-UCF-OnePager-QR.pdf

Beyond Measurement: Assessment as a Catalyst for Personalizing Learning and Improving Outcomes

Anastasia Betts, Sunil Gunderia, Diana Hughes, V. Elizabeth Owen, and Hee Jin Bang

Abstract

Despite decades of effort, summative assessments (e.g., NAEP and state standardized tests) continue to highlight persistent challenges in learning, particularly in mathematics and reading. While conventional assessments provide insights for system-level decision-making, they are often utilized as final benchmarks and can fail to address the individual ongoing needs of learners: timely, actionable feedback that directly supports student growth. This chapter introduces the Personalized Mastery Learning Ecosystem (PMLE), an adaptive, learner-centered digital system designed to address these gaps. Grounded in key learning theories such as Bloom's Mastery Learning, Vygotsky's Zone of Proximal Development, and Evidence-Centered Design, the PMLE integrates real-time formative assessments and personalized feedback to guide each learner's journey. We present a detailed worked example of the PMLE in practice, demonstrating how the system personalizes instruction, adapts to individual learner needs, and provides data-driven recommendations to educators and caregivers. The PMLE continuously adjusts to learner performance, offering tailored scaffolding to support skill development, maintain student motivation, and improve engagement. Additionally, this chapter references evidence from over twenty-five ESSA-aligned studies, including experimental and correlational studies, showcasing the significant learning gains, improved student confidence, and enhanced motivation achieved through My Math Academy® and My Reading Academy®, which embody Assessment in Service of Learning principles, particularly in supporting personalized learning and fostering equity. The chapter concludes by emphasizing the potential of the PMLE as a scalable, vertically integrated solution for modern educational challenges, offering a model for embedding formative assessment and adaptive learning into the heart of teaching and learning practices.

Introduction

Human capital—the collective knowledge, skills, and abilities of individuals—is the cornerstone of any economy's productivity and growth. A society's investment in its human capital is largely reflected in its public education system. However, recent data from assessments like the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP) paint a sobering picture of stagnant, and even declining performance in U.S. education, particularly in mathematics and reading (NAEP, n.d.; OECD, 2023; U.S. Department of Education, 2025). The 2022 PISA scores reveal a significant decline in U.S. students' math performance, while NAEP results continue to show that less than one-third of U.S. eighth graders are proficient in reading, with even steeper declines among low-income students (OECD, 2023; U.S. Department of Education, 2025).

The economic implications of these educational challenges are staggering. Stanford economist Eric Hanushek estimates that the COVID cohort of students may face a "lifetime tax" of 6% lower career earnings, potentially costing the U.S. economy $28 trillion on a present-value basis (Gunderia, 2024). This looming crisis in human capital formation necessitates urgent innovation in our approach to education, particularly in light of the rapid transformation of our economy and workforce demands driven by artificial intelligence.

While assessments like PISA and NAEP offer valuable insights into the performance of educational systems, a significant limitation of these and similar summative assessments is their failure to deliver the detailed, practical feedback necessary for teachers to enhance daily classroom instruction (e.g., Ismail et al., 2022). This limitation stems partly from the long-standing focus on the science of measurement in education, which, despite its advancements, has not been effective in informing and improving teaching and learning practices (The Gordon Commission, 2013). In contrast, more innovative approaches focus directly on improving student learning by giving both students and instructors increasingly detailed feedback on knowledge, skills, and attributes in contexts that mirror real-world applications (Behrens & DiCerbo, 2013).

Assessment in Service of Learning (AISL), as envisioned by the Handbook on Assessment in the Service of Learning, represents a paradigm shift in the role of assessment. Unlike traditional assessments that focus on measuring outcomes, AISL emphasizes diagnostic and formative assessments that inform instructional

practices and support individual student growth (The Gordon Commission, 2013). By focusing on assessments as an integral part of the learning process, AISL addresses the challenge of learner variability, ensuring that assessments adapt to the needs of diverse learners. To that end, AISL has developed a set of principles that exemplify best practices for the design of assessments in the service of learning (Table 1), which will be referred to throughout this chapter.

This chapter presents our decade-long work on developing a patented mastery-based, personalized learning system (Dohring et al., 2019, 2021, 2022) and provides an illustrative example of how learning systems can be thoughtfully designed to embody the core principles of AISL. The work presented here was carried out at Age of Learning, an international edtech company dedicated to improving learning outcomes through innovative technologies. At the core of this effort is the Personalized Mastery Learning Ecosystem (PMLE), a dynamic, adaptive digital learning system designed to provide ongoing formative assessment, personalized learning pathways, and targeted instruction. By embedding assessment directly into the learning process, the PMLE empowers an integrated and responsive approach to teaching and learning, offering a solution to the limitations of traditional assessments.

Built on patented technologies, the PMLE leverages data science, learning science and game-based assessments to adapt learning to each student's needs (Betts, 2019; Betts, Thai, & Gunderia, 2021; Thai, Betts, & Gunderia, 2022; Thai et al., 2022). These assessments provide actionable feedback to students, teachers, and families, which is essential for maintaining student motivation, engagement, and progress toward mastery.

This chapter illustrates how the PMLE reimagines assessment to support personalized learning. We begin by outlining the theoretical foundations of the PMLE, followed by a detailed worked example of the system in action, and conclude with a discussion of the evidence supporting its effectiveness. Our research, validated through over 25 ESSA-aligned studies, demonstrates that solutions built using the PMLE methodology have accelerated learning and increased student engagement (Age of Learning, 2023). These programs have proven to be particularly effective in fostering equity and ensuring that all students have the opportunity to reach their full learning potential.

By focusing on real-time feedback and adaptive learning pathways, the PMLE provides a forward-thinking model for assessment in the service of learning—one that aligns with the urgent need for innovation in today's rapidly changing world.

Table 1.
The AISL Principles are included here for convenience, as they are referenced throughout the remainder of this chapter.

Principles for Assessment in the Service of Learning	
AISL Principle 1	Assessment **transparency** provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.
AISL Principle 2	Assessment **focus** is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.
AISL Principle 3	Assessment design supports learners' **processes**, such as motivation, attention, engagement, effort, and metacognition.
AISL Principle 4	Assessments model the structure of **expectations** and **desired learning** over time.
AISL Principle 5	**Feedback**, adaptation, and other relevant instruction should be linked to assessment experiences.
AISL Principle 6	Assessment **equity** requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.
AISL Principle 7	**Assessment quality and validity** should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Theory of Change

The Theory of Change underlying the PMLE begins with the premise that effective learning systems must adapt to the unique needs of each learner, reflecting the variability inherent in how students learn and progress. Barbara Pape (2018) and her colleagues at Digital Promise emphasize that learner variability is the norm, not the exception, encompassing a wide range of cognitive, social-emotional, and environmental factors that shape student learning experiences. Traditional education systems often fail to accommodate this variability, relying instead on a one-size-fits-all model grounded in the myth of the "average" learner—a fallacy, Todd Rose (2016) critiques in *The End of Average*.

The PMLE rejects the notion of "average" and instead embraces the complexity of learner variability through personalized, adaptive environments (Fig. 1). These environments are characterized by real-time formative assessments, dynamic learning activities, continuous feedback, and tailored scaffolding that work collectively to promote optimal learning within each student's Zone of Proximal Development (ZPD; Betts et al., 2024; Vygotsky, 1978). These assessments further enable the creation of personalized learning pathways designed to address gaps in foundational knowledge while simultaneously fostering motivation through game-based contexts that celebrate individual mastery-based progress rather than comparative benchmarks.

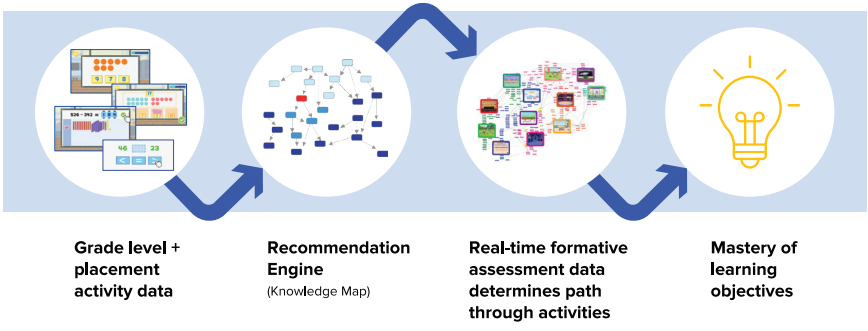


Figure 1.

The student experience begins with a set of placement activities that approximate the student's current ZPD. From there, the Recommendation Engine selects activities appropriate for the student. Data from those activities is used to determine the next set of recommendations. As the student progresses, they master learning objectives and are recommended new ones.

In the broader design of the PMLE, educators and caregivers play a pivotal role. Equipped with real-time data on each learner's knowledge, skills, and progress, they can offer encouragement, transfer activities, and targeted interventions that amplify the system's impact. Research study findings on the PMLE include improved student academic outcomes and increased learner engagement and confidence. Ongoing research suggests that over time, the PMLE's approach has the potential to reduce educational disparities, enhance instructional effectiveness, and provide broader societal benefits, including strengthened human capital and equity in educational opportunities (Age of Learning, 2023).

Theoretical Framework

The PMLE is grounded in robust educational theory, ensuring that instruction and assessment are aligned with how students naturally learn and develop. It is informed by several foundational theories from cognitive science and educational psychology, which provide a comprehensive understanding of the processes that underlie personalized learning. Central to this framework is the recognition of learner variability, which posits that every learner possesses unique combinations of prior knowledge, skills, and needs for future learning (Rose, 2016). This variability requires a theoretical framework for the design of complex adaptive learning environments that accommodate and support learners as they progress through their individual learning journeys.

The design of the PMLE deliberately draws upon and synthesizes multiple complementary frameworks, rather than relying on a single theoretical approach. This integration accommodates the multifaceted nature of personalized learning, particularly when situated in "smart learning" environments. Each theoretical component contributes unique insights including understanding how learners progress toward mastery, mapping knowledge structures, supporting optimal cognitive development, designing aligned assessments, and leveraging learning analytics. Together, these frameworks provide a comprehensive foundation for creating an adaptive learning environment that can effectively respond to individual learner differences while maintaining pedagogical rigor and evidence-based practice. In the following sections, we examine each of these in detail, exploring how they individually and collectively inform the PMLE's design and implementation.

Theory of Mastery Learning

In his theory of Mastery Learning, Bloom (1968) argued that most students can achieve a high level of mastery if given the right conditions—time, appropriate instruction, and formative feedback. Central to this theory is the idea that learning difficulties often arise when learners are not sufficiently prepared for new material due to gaps or misunderstandings in foundational knowledge (Bloom, 1984). According to Bloom, the learning process can be optimized by ensuring that students first master prerequisite knowledge before moving on to more complex topics. Mastery learning emphasizes a structured, step-by-step approach where feedback and corrective actions are integrated into instruction, allowing learners to progress only when they have fully grasped earlier material. Bloom's theory is vital for understanding how learners differ in their progress, as it highlights the need for systems that can accurately assess where each learner is on their learning trajectory and provide personalized interventions to address gaps and misunderstandings before moving forward.

Theory of Objects of Change

Building on his work in mastery learning, Bloom (1984) identified four critical "objects of change" that must be addressed for effective learning: the student, the teacher, the materials, and the learning environment. This framework emphasizes that successful educational interventions must consider and support all four elements in concert. According to Bloom, focusing on any single object of change in isolation is insufficient; rather, meaningful educational improvement requires coordinated attention to how each object interacts with and supports the others. The student must be supported through appropriate instruction and feedback; the teacher must be equipped with necessary tools and insights; the materials must be high-quality and responsive to learning needs; and the learning environment—including both physical spaces and social contexts like family and peer interactions—must be enriched to support learning goals. This holistic framework provides crucial guidance for designing comprehensive educational systems that can effectively support learning at scale.

Knowledge Space Theory

To effectively personalize learning, it is also necessary to model the structure of knowledge within a given domain. Knowledge Space Theory (KST), developed by Falmagne and Doignon (1999), offers a framework for representing the relationships between different concepts, principles, and skills within a subject area. Knowledge modeling (KM), a key aspect of KST, also provides a detailed map of the possible learning paths that a student might take through content, depending on their existing knowledge state. By modeling these intricate relationships, KST allows for the identification of each learner's current knowledge state—the collection of concepts they have mastered—and what they are ready to learn next. The application of KST, and more specifically the process of knowledge modeling and mapping learner knowledge states against that knowledge model, is central to the PMLE's ability to offer targeted instruction, ensuring that each learner receives content that aligns with their current level of understanding while avoiding material that is too advanced or redundant.

Zone of Proximal Development

Vygotsky's (1978) theory of the zones of development is another cornerstone of the PMLE's theoretical foundation. Vygotsky posited that while there are three zones of development (e.g., Zone of Actual Development, Zone of Insurmountable Difficulty, etc.), it is the Zone of Proximal Development (ZPD) where learning is most efficient. In the ZPD, learners are challenged to extend their abilities with the support of a more knowledgeable other (MKO)—whether that be a teacher, peer, or intelligent system.

Vygotsky's ZPD theory emphasizes the importance of scaffolding—providing the right amount of support at the right time to help learners progress. This theory underpins the PMLE's approach of using dynamic scaffolding in its instructional activities, where learners receive assistance when needed–as they would when working with an MKO–but are encouraged to achieve independence over time. Our leveraging of the learner's ZPD recognizes that even when learners may have achieved the same or similar levels of mastery, their ability to stretch into more complex content is variable. As such, the PMLE deploys adaptive features that can explore and exploit this ZPD "elasticity" to adjust each individual's personal rate of progress and stretch each learner toward unique learning goals (Betts et al., 2024).

Evidence-Centered Design

Evidence-Centered Design (ECD), developed by Mislevy and colleagues (2003), provides a structured framework for creating assessment tasks that yield valid evidence of specific learning outcomes. At its core, ECD ensures that successful task completion genuinely demonstrates mastery of intended learning objectives. The PMLE builds on this foundation by refining learning trajectories and grounding progress in credible evidence of the learner's growth.

The PMLE implements ECD principles through the design of assessments that are seamlessly integrated into the learning process, transforming assessment from a summative tool into an ongoing, dynamic component of instruction. Each task is intentionally crafted to generate clear, interpretable evidence of a learner's knowledge, skills, and abilities by eliminating construct-irrelevant paths to success. This precise measurement ensures that when a student succeeds at a task, that success truly reflects their understanding of the target objective.

This careful application of ECD principles generates a continuous stream of actionable data about what learners know and what they are most ready to learn next. The real-time evidence drives the PMLE's personalization engine, enabling the system to deliver precisely targeted instruction based on valid evidence of each learner's current understanding. This tight alignment between learning objectives, assessment tasks, and instructional decisions creates a responsive learning environment that consistently advances student mastery.

Educational Data Mining (EDM)

Educational Data Mining (EDM) is a critical component of modern adaptive learning systems, enabling data-driven personalization of instruction through the analysis of learner interactions, errors, response times, and patterns of success (e.g., Romero & Ventura, 2010). Particularly in game-based learning environments, discovery through experimentation in play is an implied norm of games (Salen & Zimmerman, 2004), resulting in rich event-stream data logs that can reveal actions and pathways that unfold as students engage in behaviors like subversive play and productive failure (e.g., Owen et al., 2019). EDM is a discipline that can help us understand student choices like these in the complex digital systems of educational games (Owen & Baker, 2019) – providing a broad range of methods for the organization, research, and analysis of big data in complex learning environments.

By leveraging machine learning and statistical techniques, EDM uncovers trends in student behavior, allowing systems to predict future needs and make real-time adjustments to content, scaffolding, and feedback. This continuous cycle of data collection and analysis supports more precise, evidence-based adaptation, ensuring that learning pathways evolve dynamically in response to each learner's progress. By integrating EDM into adaptive learning environments, instructional systems can move beyond static personalization, instead offering a responsive, data-informed approach that optimizes learning efficiency and supports individual growth.

A Learning Sciences Framework for Design

The integration of these theories reflects core principles in the learning sciences about how people learn and develop expertise. By weaving together Bloom's insights on mastery progression, Knowledge Space Theory's systematic mapping of relationships between subject-domain concepts and skills, Vygotsky's understanding of scaffolded development, Evidence-Centered Design's approach to meaningful assessment, and Educational Data Mining's capacity for deriving actionable insights from patterns in the data, we create a theoretically grounded system that responds to the fundamentally unique learning processes of every individual learner's, *at scale*.

This comprehensive framework acknowledges that learning is not just about content delivery, but involves complex interactions between cognitive development, knowledge construction, social support, and individual differences. This integration enables the system to provide appropriate challenges, targeted support, and meaningful feedback—key elements that learning sciences research has shown to be crucial for effective learning. This robust theoretical grounding sets the stage for our discussion of Assessment in Service of Learning (AISL) in the next section, where we explore how these learning sciences principles are operationalized in practice.

Worked Example: Personalized Mastery Learning Ecosystem

The theories and principles discussed in the previous section are operationalized in the Personalized Mastery Learning Ecosystem (PMLE) through two flagship programs: *My Math Academy*® and *My Reading Academy*®. These programs enhance early math and literacy learning through game-based, adaptive instruction that responds to each learner's unique needs. For example, My Math Academy focuses on building strong foundations in early math concepts and skills while My Reading Academy develops essential literacy skills including decoding, vocabulary, and comprehension. Using the PMLE, both programs seamlessly integrate assessment into engaging learning activities, creating an experience that continuously adapts to learner progress while maintaining high levels of engagement.

These implementations of the PMLE have demonstrated significant success in improving early learning outcomes across diverse educational settings, as numerous ESSA-aligned studies have shown their effectiveness in increasing student achievement, engagement, and confidence (Age of Learning, 2023). In the following sections, we examine the key components that enable this success, exploring how each element of the PMLE works together to create a cohesive, personalized learning experience.

Key Components of the Personalized Mastery Learning Ecosystem

The PMLE is composed of several interconnected components, each designed to support personalized instruction and real-time assessment. These components—discussed in the sections that follow—work together to provide a seamless, adaptive learning experience for students, while offering educators and caregivers critical insights to guide instruction and support.

Knowledge Map

Applying the knowledge modeling aspect of KST, the "Knowledge Map" forms the backbone of the PMLE, outlining the structure of the learning objectives within a specific domain, such as mathematics or reading (Fig. 2). This map organizes concepts and skills into a coherent, hierarchical framework, illustrating the relationships between them—such as prerequisite knowledge, sequential learning, or parallel development. By mapping out the myriad pathways a learner might take, the system can pinpoint what each student already knows, what they are

most ready to learn next, and predict where they might encounter challenges. The Knowledge Map models the structure of expectations and desired learning over time (AISL Principle 5), and ensures that assessment is transparent, providing clarity on what students are expected to learn and how their progress will be measured (AISL Principle 1). In the PMLE, the next steps in the learning process are clear, creating a roadmap for individual student growth.

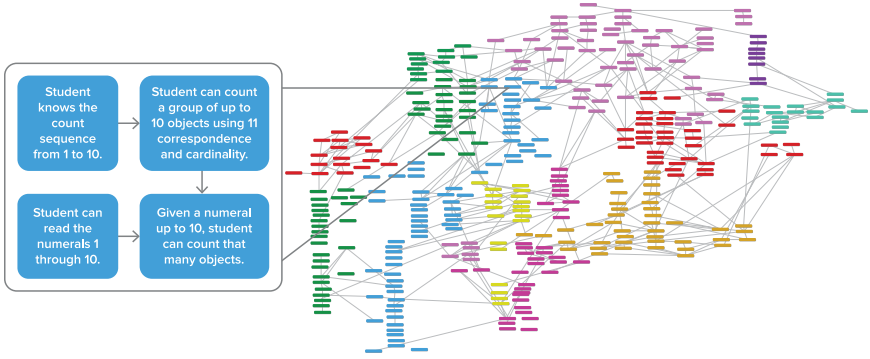


Figure 2

My Math Academy knowledge map overview of PreK-2 number sense and operations. Each block represents a learning objective.

The Knowledge Map also serves as the essential foundation for implementing Vygotsky's Zone of Proximal Development (ZPD) theory within the PMLE. By comprehensively mapping the relationships between learning objectives, it creates a structured space where the system can identify each learner's optimal learning zone. This mapping is crucial for the PMLE's core mission of maximizing learning efficiency—ensuring students engage with content that challenges them appropriately. The Knowledge Map essentially charts the complete terrain of possible learning pathways, allowing the system to pinpoint where each student's ZPD lies and adapt instruction accordingly.

Learning Activities

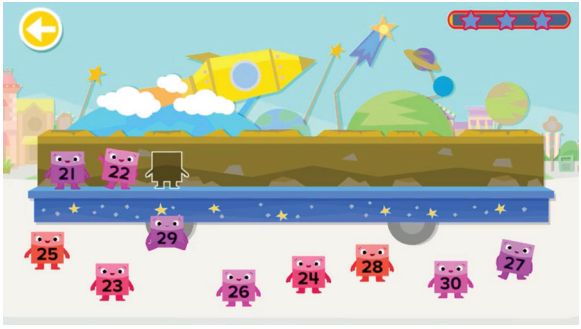
Within the PMLE, Learning Activities are the primary method through which students engage with instructional content. These activities are divided into two main types: Direct Instruction and Scaffolded Assessment. Direct Instruction involves explicit teaching of concepts, often through interactive games, videos, or demonstrations. Scaffolded Assessment provides opportunities for students to practice and apply their knowledge while receiving immediate feedback and support when needed, similar to what they would receive with a live MKO.

These activities are designed to ensure that learners remain within their ZPD, where tasks are challenging but achievable. As students interact with the system, embedded assessment features monitor their performance and adapt the level of support accordingly. When students struggle, the system offers scaffolding, such as modeling based on prior knowledge, step-by-step breakdowns, or additional practice opportunities, ensuring that learning continues in a supportive environment (Figure 3). Teachable moments are strategically leveraged throughout the learning experience to optimize growth.

This design component stresses the need for assessment features that support the learner's motivation, engagement, and self-regulation (AISL, Principle 4). By embedding formative assessment within the learning activities, the PMLE provides real-time feedback that guides the learner's process and keeps them motivated to achieve mastery. The ability to adapt based on performance also promotes engagement, as students are continually challenged at their "just-right" level.

The Shapeys Line Up for a Parade: Count Sequences in My Math Academy

Learner is asked to build one group of ten numbers in the counting sequence between 21 and 100.

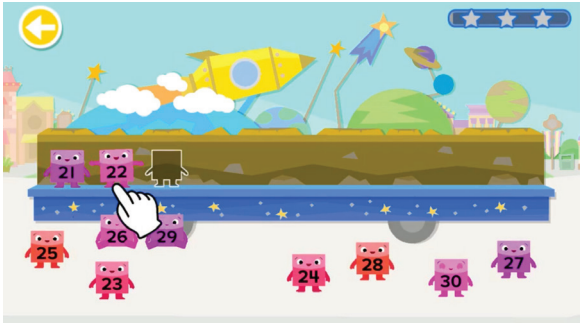


Phase 0

The student hears the following voiceover: "The Shapeys are having a parade! Count forward from 21 to put the Shapeys on the float." The first two Shapeys count off - "Twenty-one" "Twenty-two!" The narrator follows with "What number comes after twenty-two?" The game is now open for student input (e.g., dragging Shapeys)

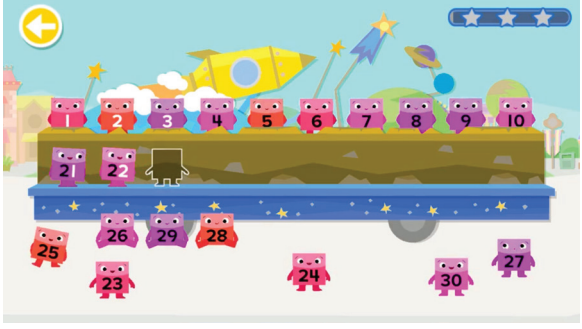
Phase 1

The student has answered incorrectly. Their incorrect Shapey choice sits down to indicate it cannot be chosen again. The student hears, "That's not the next number. What number comes after 22?" The game is now open for input.



Phase 2

The student has answered incorrectly. Their incorrect Shapey choice also sits down to indicate it cannot be chosen again. The student hears, "Uh oh! That's not the next number. Try counting forward from 21 to find what number comes next." A "helping hand" guides the student to tap each of the preset Shapeys in succession - as they tap, the student hears the Shapeys say "Twenty-one!" and "Twenty-two!" The game is now open for input.



Phase 3

The student has answered incorrectly. Their incorrect Shapey choice also sits down to indicate it cannot be chosen again. The student hears "Uh oh! That's not the next number." A row of Shapeys pops up from behind the float, displaying numbers 1–10. The narrator says "Let's look at the number pattern. When we start counting from one, we count..." and the Shapeys count off "one, two, three!" Narrator follows with, "So when we start counting from 21, we count..." and the Shapeys count off "21, 22!" Narrator ends with "What number comes after 22?" The game is now open for input.

Phase 4
The student has answered incorrectly. Their incorrect Shapey choice also sits down to indicate it cannot be chosen again. The student hears "Uh oh! That's not the next number." A row of Shapeys pops up from behind the float, displaying numbers 1–10. The narrator says "Let's look at the number pattern. When we start counting from one, we count..." and the Shapeys count off "one, two, three!" Narrator follows with, "So when we start counting from 21, we count..." and the Shapeys count off "21, 22!" Narrator ends with "What number comes after 22?" The "helping hand" appears and guides the player to drag in the correct Shapey, who shouts out "Twenty-three!" as it is dragged.

Figure 3.
A sample of a Scaffolded Assessment activity, showing the unscaffolded formative assessment task, and then increasing layers of scaffolding corresponding to repeated errors by the student.

Personalization Engine

The Personalization Engine is the heart of the PMLE, continuously analyzing real-time data from Learning Activities to guide each student's unique pathway through the Knowledge Map. As students interact with the system—whether succeeding or struggling—the engine determines their optimal next steps, from advancing to new concepts to reviewing previous material or adjusting scaffolding levels.

Drawing on the concept of ZPD elasticity, the engine recognizes that learners at similar knowledge levels may vary significantly in their capacity to stretch toward more advanced content (Betts et. al., 2024). By measuring and responding to each learner's individual growth potential within their ZPD, the system dynamically adjusts instruction to maintain an optimal level of challenge, particularly benefiting students at risk of falling behind.

The engine operationalizes Vygotsky's theories by synthesizing two critical data streams: the Knowledge Map's comprehensive view of possible learning pathways and real-time formative data from learning activities. This synthesis enables the engine to precisely locate each student's ZPD and deliver appropriately challenging activities that can be completed with the system acting as a proxy for an MKO.

This sophisticated personalization upholds key Assessment in Service of Learning principles by ensuring equitable access to learning opportunities (AISL Principle 3) while maintaining transparency about learning expectations, student progress, and next steps (AISL Principle 1). The engine's continuous adaptation ensures that every learner receives instruction tailored to their current knowledge state and growth potential, supporting their individual journey toward mastery.

Educator and Caregiver Centers

The Educator and Caregiver Centers are core components of the PMLE designed to provide actionable insights into student progress. Educators receive real-time performance data along with targeted recommendations for instructional adjustments, including suggested student groupings based on learning readiness and specific intervention or enrichment activities. The Caregiver Center equips parents and caregivers with detailed progress reports and targeted activities to support their child's learning journey.

These centers fulfill Bloom's (1984) Objects of Change framework by addressing all four key elements:

- Supporting the **Student** through a personalized, encouraging learning environment focused on their existing knowledge and ZPD
- Providing high-quality, rigorous, and responsive **Materials**
- Assisting the **Teacher** with data and resources for effective differentiated instruction
- Enriching the **Learning Environment** by empowering **Family Members** with tools and insights to support their child's learning

By providing precise data about each student's ZPD along with specific support recommendations, these centers enable teachers and caregivers to elevate their ability to serve as MKO in Vygotsky's framework, whether in the classroom or at

home. Moreover, the centers ensure transparency around learner progress while delivering actionable feedback and recommendations to all stakeholders (AISL Principle 7). This coordinated approach creates a coherent learning experience across different environments, with data-driven insights enabling timely, informed decisions that strengthen the connection between assessment, instruction, and learning outcomes.

Figure 4.

Educator Center display of students in a class who have completed skills (green), are making progress in skills (blue), or need support in skills (orange).

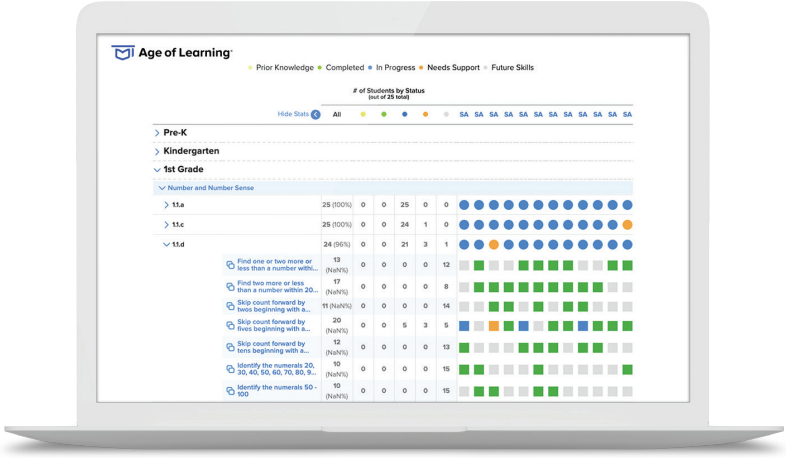


Figure 5.

Educator Center display of groupings of students along with a video explaining what students are working on and why it is important.

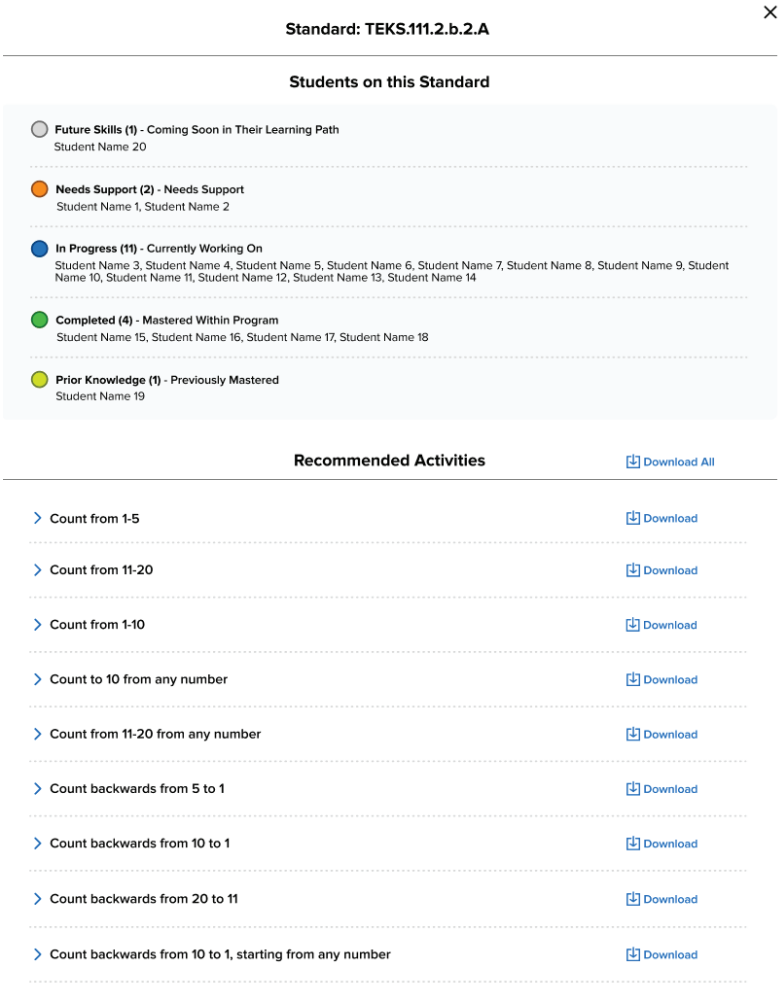


Figure 6.

Group Activity Recommendation available in the Educator Center to support students working on counting to 10

Back to Skill/Concept

Count from 1-10

Recommended Activity

Instruction Types: One-on-One Activity | **Group Activity**

Print/Download

Counting Coaches (1–10)

(Estimated Time: 15 mins.)

Purpose To identify mistakes in counting and describe how to count to 10.

Materials

Students are often motivated to help someone else learn. Puppets serve as a good "learner" when using this method of instruction. If a puppet is not available, you can pretend that you "forgot" how to count.

- puppet (optional)
- Number Line 1–10 (optional) [PDF]

Vocabulary

none

Activity

Tell students that they will help someone learn to count.

Start by asking students to count together from 1 to 10.

Ask: *When are times that you or someone else has counted to 10?*

> Have students share real-life examples of counting, such as counting objects or when playing a game like hide-and-go-seek. Discuss why it is important to count the numbers in the same order for each example.

Explain that you (or a puppet) need help counting to 10. Ask students to listen carefully for any mistakes, but not to shout out when they hear it. Complete the first example as a group. **Say:** *one, two, three, four, five, seven, six, eight, nine, ten.* [six and seven counted out of order]

Ask: *Did I make any mistakes?*

> Students should notice that the numbers six and seven were counted out of order. Have students explain their reasoning. For example: *After the number five, you count six. After the number six, you count seven.*

Continue counting to 10 while making mistakes. Ask partners to discuss the mistake before sharing with the group. Sample sequences:

- *one, two, three, four, five, six, seven, nine, ten* [skipped eight]
- *one, two, three, four, five, five, six, seven, eight, nine, ten* [five counted twice]
- *one, two, three, four, five, six, eight, ten* [skipped seven and nine]

Finish by counting from 1 to 10 correctly. When no mistakes are found, thank students for helping you (or the puppet) count to 10.

Exit Ticket

Choose a number from 5–9. **Ask:** *What number comes after _____ when counting?*

Success Strategies

Teacher

Using a number line can support students in knowing if a number was skipped, omitted, or repeated when listening to the count.

Student

Discuss with students that making mistakes is how we learn. Ask students to share a time when they learned from making mistakes.

English Learner Support

Restate and Repeat—Repeating the count sequence at a slower pace can be helpful by providing additional processing time.

Hypothetical Case Study: Maya's Learning Journey

To illustrate how the PMLE functions in practice, consider the following hypothetical journey of a young learner named Maya, who has been struggling with foundational math concepts, particularly multiplication. Maya's experience demonstrates how the PMLE adapts to her individual needs while embodying the key principles of Assessment in Service of Learning.

Initial Assessment and Placement

Maya begins her journey with an initial diagnostic assessment embedded within the PMLE. The system evaluates her current knowledge and identifies that she has unfinished learning related to her understanding of multiplication basics, which are crucial for her progression in math. Based on this, Maya is placed in an evolving learning pathway that starts with reviewing prerequisite skills like skip counting and repeated addition. This assessment models Maya's learning trajectory by identifying her knowledge state in relation to the PMLE's knowledge map (e.g., what Maya knows and has already mastered, etc.), and what she is most ready to learn next (AISL, Principle 5).

Learning Activities and Real-Time Feedback

Maya engages with the system through interactive games and activities designed for Direct Instruction and Scaffolded Assessment. In one activity, Maya works on a game where she must group objects to understand the concept of multiplication as repeated addition. Initially, she struggles, and the system responds by offering scaffolds such as visual aids and reminders of previous strategies to guide her. She receives real-time feedback that encourages her persistence, and the system monitors her progress, adjusting the difficulty level based on her performance. As Maya begins to grasp the concept, the scaffolds are gradually removed, allowing her to complete the task independently (i.e., demonstrating mastery). By monitoring Maya's interactions and ensuring that she remains in her ZPD, the system ensures she is challenged but not frustrated, maintaining her motivation to learn. Through these mechanisms, the PMLE not only assesses Maya's moment-to-moment learning, but also enhances her motivation, engagement, and metacognition (AISL, Principle 4).

Personalized Pathway and Equity

After mastering multiplication basics, Maya is ready to move forward. The Personalization Engine evaluates her success and adjusts her pathway to introduce more complex multiplication problems. However, when Maya encounters difficulty again, the system recognizes this and automatically reintroduces scaffolds, ensuring that she continues to receive the support she needs without feeling frustrated or discouraged. This process ensures equity by adapting instruction to meet Maya's unique moment-to-moment learning needs, providing her with the personalized resources necessary for continued success (AISL, Principle 3).

Educator and Caregiver Involvement

While Maya progresses through the system, her teacher and caregivers are kept informed through the Educator and Caregiver Centers. Her teacher accesses real-time data from the Educator Center, which shows that Maya is struggling with the transition from conceptual understanding to applying multiplication facts. The system recommends targeted group activities and one-on-one interventions to help Maya reinforce these skills in the classroom.

At home, Maya's caregivers receive similar insights through the Caregiver Center, along with suggestions for simple math games they can play together to reinforce multiplication concepts. These real-time, actionable insights ensure that all stakeholders—educators, caregivers, and Maya herself—have the information they need to make informed decisions and support her learning journey (AISL, Principle 7).

Continuous Progression

Also a result of her learning journey, Maya has not only mastered multiplication but has gained confidence in her math abilities. The PMLE continues to push her forward, introducing division as the next logical step in her learning path. Throughout her experience, Maya has benefited from a system that adapts to her learning needs, provides timely feedback, and engages all stakeholders in supporting her progress. In doing so, the PMLE exemplifies how assessments can be reimagined to foster personalized, equitable learning experiences that go beyond mere evaluation and become integral to the learning process.

Maya's Journey: The Learning Sciences at Work

Maya's journey through the PMLE exemplifies how multiple learning science theories work together to create an effective learning system. Bloom's (1968) theory of mastery learning ensures complete understanding of foundational concepts before advancement, while his Objects of Change framework (1984) coordinates support across four essential elements: providing Maya with personalized instruction, equipping her teacher with actionable data, delivering responsive learning materials, and enriching her learning environment through family engagement.

Vygotsky's (1978) zones of development, specifically his theory of the ZPD, shapes both the dynamic scaffolding Maya receives and the selection of her next learning activities. The system serves as a virtual More Knowledgeable Other MKO, providing just-in-time support while also strategically selecting activities that stretch her capabilities, leveraging the elasticity of her ZPD to optimize learning gains. This theoretical foundation integrates seamlessly with Knowledge Space Theory (Falmagne & Doignon, 1999), which guides the mapping and sequencing of Maya's learning progression through the domain content, ensuring that the system can identify optimal opportunities to stretch her understanding while maintaining appropriate support.

Evidence-Centered Design principles (Mislevy et al., 2003) provide the foundation for how the system gathers valid evidence of Maya's understanding. Each task she encounters is intentionally designed to eliminate construct-irrelevant paths to success, ensuring that her achievements genuinely reflect her mastery of targeted concepts. This precise assessment framework works in tandem with Educational Data Mining (EDM), which analyzes patterns in Maya's interactions, response times, error types, and learning progressions to allow for system iteration over time. This continuous analysis not only predicts her immediate learning needs but also reveals broader patterns in her learning approach, allowing the system to optimize the timing, type, and level of support she receives. Together, these frameworks ensure that every interaction generates meaningful data that drives increasingly precise personalization of Maya's learning experience.

This integration of learning sciences creates a theoretically grounded system where each framework serves a distinct yet complementary purpose: Bloom's theories ensure mastery and comprehensive support, Vygotsky's ZPD and Knowledge

Space Theory guide optimal progression, while ECD and EDM enable precise assessment and personalization. Together, these create a learning experience that exemplifies Assessment in Service of Learning principles by supporting motivation through appropriate challenges, providing transparent feedback through valid assessments, and ensuring equitable access through personalized support.

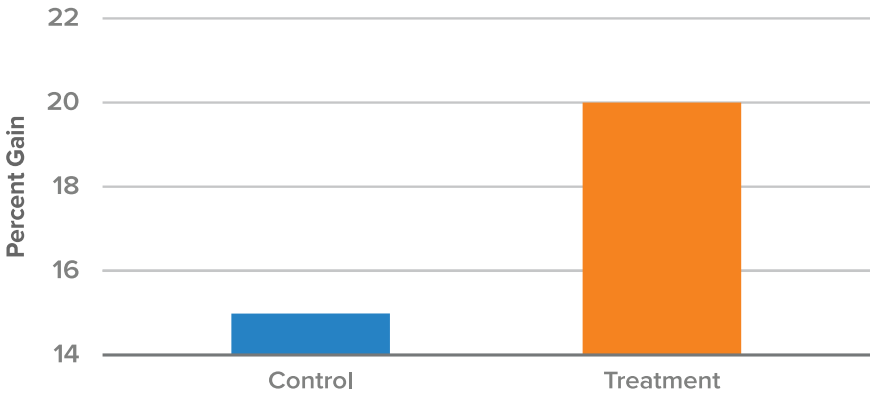
Evidence Of Efficacy

The PMLE, operationalized through My Math Academy® and My Reading Academy®, has undergone extensive evaluation across diverse educational contexts, consistently demonstrating meaningful gains in early math and literacy achievement. To date, programs have been the subject of 27 ESSA-aligned studies each—spanning randomized controlled trials (RCTs), quasi-experimental designs, and correlational analyses—with total sample sizes of 19,494 for My Math Academy® and 25,488 for My Reading Academy®. The breadth of these studies is noteworthy, encompassing Title I schools, urban, suburban, and rural communities, and serving populations that include high proportions of Hispanic/Emergent Multilingual Learners, students from low-income families, and students with Individualized Education Plans. Such robust and comprehensive research underscores the adaptability and efficacy of the PMLE framework in addressing critical educational equity challenges (Age of Learning, 2025).

My Math Academy®

My Math Academy® has proven effective in helping young learners—including students with individualized education plans—build foundational math skills. One significant randomized controlled trial, conducted in California in partnership with WestEd, evaluated over 400 kindergarten and transitional-kindergarten students across Title I public schools. The students were randomly assigned to either use My Math Academy™ or continue with traditional instruction for 12–14 weeks. Results showed that students in the My Math Academy® treatment group exhibited statistically significant improvements in early mathematics skills compared to the control group, with an effect size of 0.23, $p < 0.05$ (Fig. 7; Thai, Bang, & Li, 2021).

Figure 7.
***Percent gain in TEMA-3 math scores by treatment group students who used My Math Academy* (n = 233) *and control group students who did not* (n = 195, p < .05, effect size = 0.23)**



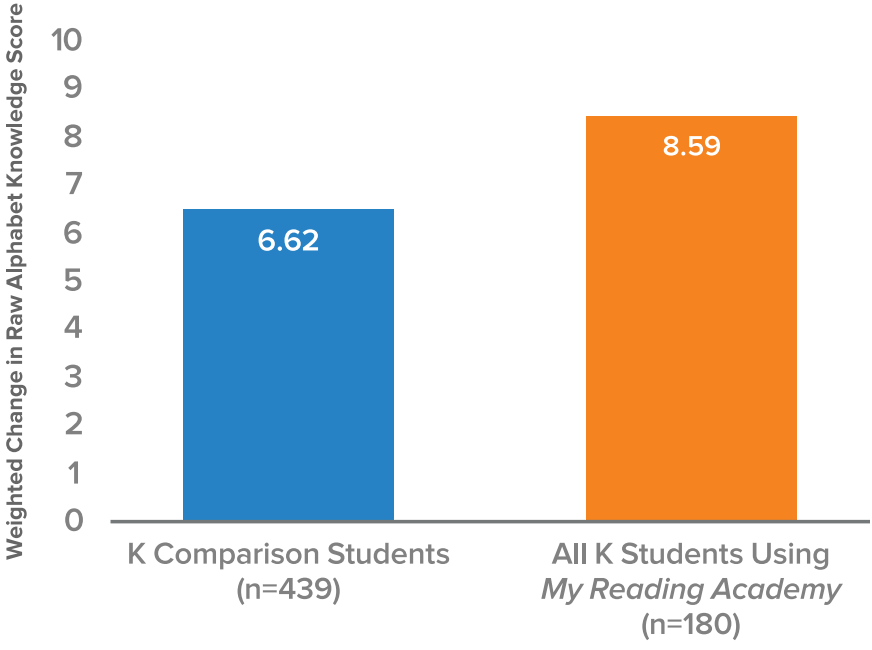
The study's findings revealed that students who used My Math Academy® demonstrated the greatest learning gains on more difficult math skills, highlighting the program's ability to scaffold instruction and support learners as they progress through increasingly challenging content. These results were independently verified by LearnPlatform by Instructure and Evidence for ESSA, confirming the program's alignment with Tier I standards for evidence-based interventions (Age of Learning, 2025).

My Reading Academy®

Similarly, My Reading Academy® has shown considerable impact in early literacy development. In the 2021–2022 school year, Age of Learning partnered with SRI International to conduct a quasi-experimental study across two states (Virginia and Texas), evaluating over 1,000 pre-kindergarten and kindergarten students. Results demonstrated that students who used My Reading Academy® outperformed their peers on key literacy metrics, such as alphabet knowledge, with an effect size of 0.29, $p < 0.05$ (Fig. 8; Bang & Siebert-Evenstone, 2025).

Figure 8.

Change in the raw Alphabet Knowledge score from fall to spring



More importantly, students who spent more time engaged with the program exhibited even greater gains. For example, kindergarten students who averaged 60 minutes of use per week achieved an effect size of 0.47 compared to their peers, reinforcing the link between usage levels and outcomes. Additionally, among pre-kindergarteners, those who mastered at least 16 alphabet knowledge skills were significantly more likely to be "on track" in reading by the end of the school year.

Impact on Teachers and Classrooms

The qualitative impact of My Math Academy® and My Reading Academy® has been equally compelling. Nine out of 10 teachers who used these programs reported that they provide a technology-rich environment that fosters student engagement and learning (Age of Learning, 2025). Educators noted the programs' ability to empower students by increasing their self-confidence and enjoyment of learning, particularly in areas where students had previously struggled. Teachers also emphasized the value of real-time data and feedback, which enabled them to tailor instruction more effectively to individual student needs.

These outcomes emphasize the importance of designing assessments that support learners' motivation, attention, and metacognition (AISL, Principle 4). The consistent reports of increased student engagement and self-confidence indicate that My Math Academy® and My Reading Academy® are successfully fostering the kind of intrinsic motivation needed for long-term academic growth. Furthermore, the real-time feedback provided to educators through these systems reinforces the need for assessments to offer actionable insights to inform instructional decisions (AISL, Principle 7).

Lessons Learned

The demonstrated efficacy of the PMLE in achieving significant learning outcomes across two distinct domains—reading and math—highlights the versatility and robustness of its framework. By addressing foundational skills in these critical areas, the PMLE shows that its principles and methodologies can be successfully adapted to varied learning objectives and contexts. This versatility is not coincidental; it is a direct result of a deliberate, evidence-driven design process that integrates research-based curricula, user-centered design, and multidisciplinary collaboration.

Key insights from our journey developing the PMLE point to the foundational role of a disciplined, evidence-based approach in creating effective educational solutions. Two core principles underpinned this success: research-based curricula, as articulated by Clements (2007) in his framework for research-driven design, and design thinking, which prioritizes user needs to ensure solutions are empathetic, impactful, and practical in diverse contexts (Brown, 2008).

Crucially, our development process reflects the complexity of human learning. No single domain of learning science provides a comprehensive roadmap; instead, innovation demands an integrative approach. To achieve this level of integration, the design and development of the PMLE relied on a multidisciplinary team that brought together experts in curriculum development, cognitive science, data science, software engineering, game development, psychometrics, behavioral science, and educational data mining. This collaborative "team sport" approach—often referred to as Learning Engineering—enabled us to balance scientific rigor with creative design, ensuring that the PMLE was both engaging for learners and effective in producing measurable outcomes.

In addition to its interdisciplinary foundation, the PMLE exemplified a continuous improvement philosophy. Robust user research ensured that the learner interface was developmentally appropriate and intuitive, while iterative cycles of testing and evaluation allowed us to optimize both the content and delivery of learning experiences. Efficacy testing was key to understanding not only whether the solutions were effective but also for whom and in what contexts. This iterative design process deepened our insights into the cognitive, social, and environmental factors that shape learning, reinforcing the value of human-centered design.

The broader implications of the PMLE extend beyond its demonstrated success in reading and math. It provides a scalable model for addressing educational equity by tailoring learning experiences to individual needs while maintaining scientific rigor. Moreover, its development process offers a replicable framework for other educational innovations, emphasizing the necessity of evidence-driven design, interdisciplinary collaboration, and a commitment to continuous refinement.

Conclusion

The design of the Personalized Mastery Learning Ecosystem represents a significant leap forward in addressing some of the most pressing challenges in modern education: learner variability, lack of personalized feedback, and the limitations of traditional assessment methods. Through its integration of real-time formative assessment, adaptive learning paths, and dynamic scaffolding, the PMLE exemplifies how assessment can evolve beyond mere measurement to become an essential driver—a catalyst—of personalized, learner-centered education.

Grounded in robust learning science principles, such as Bloom's Mastery Learning and Objects of Change, Knowledge Space Theory, Vygotsky's Zone of Proximal Development, Evidence-Centered Design, and Educational Data Mining, the PMLE has demonstrated its ability to support student growth in both math and literacy. The evidence from nearly thirty ESSA-aligned studies shows substantial gains in student learning outcomes, increased engagement, and improved motivation across diverse student populations, including those who are historically under-served. These results are not only a testament to the efficacy of the PMLE, but also a reflection of its potential to close equity gaps in education.

Aligned with the AISL principles, the PMLE emphasizes assessment transparency, equity, and actionable feedback for learners, educators, and caregivers. These principles underpin the system's design, ensuring that all stakeholders have access to the data and insights they need to make informed decisions and support student success. The PMLE also underscores the importance of motivation and engagement, offering students the right level of challenge and support to keep them engaged in their learning journey.

As we move forward in an era of rapid technological change, systems like the PMLE offer a vision for how assessment can be reimagined to foster not only academic achievement, but also a deeper, more personalized connection to

learning. By embedding formative assessment directly into the learning process and offering real-time, tailored feedback, the PMLE aligns perfectly with the overarching goals of the AISL movement: to ensure that assessments stimulate, support, and develop learning.

In conclusion, to address the pressing challenges in education, it is imperative to embrace the transformative principles of the PMLE and the AISL movement. Through the development of thoughtfully designed adaptive technologies, we can redefine assessment—not as a mere measure of progress, but as a vital tool for aligning learning resources to the unique needs of every student. This approach paves the way for fostering inclusivity, advancing evidence-based practices, and building the human capital essential for creating a thriving, equitable society.

References

Age of Learning. (2023). *Research-driven approach.* https://www.ageoflearning.com/research

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Bang, H. J., & Siebert-Evenstone, A. (2025). Evaluation of a Personalized Game-Based Learning app for Developing Young Children's Reading Skills. *Journal of Educational Technology Systems, 54*(1), 88-118. https://doi.org/10.1177/00472395251337095

Behrens, J. T., & DiCerbo, K. E. (2014). Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment. *Teachers College Record*, *116*(11), 1–22.

Betts, A. (2019). Mastery learning in early childhood mathematics through adaptive technologies. In IAFOR (Ed.), *The IAFOR International Conference on Education–Hawaii 2019 Official Conference Proceedings.* Paper presented at the IAFOR International Conference on Education: Independence and Interdependence, Hawaii (pp. 51–63). Japan: The International Academic Forum.

Betts, A., Hughes, D., Plache, L., & Smith, K. (2024). Stretching the Zone of Proximal Development: Accelerating learning through ZPD elasticity. *The IAFOR International Conference on Education–Hawaii 2024 Official Conference Proceedings*.

Betts, A., Thai, K. P., & Gunderia, S. (2021). Personalized Mastery Learning Ecosystems (PMLE): Using Bloom's four objects of change to drive learning in Adaptive Instructional Systems (AISs). In *Human-Computer Interaction International 2021 Conference Proceedings, Lecture Notes in Computer Science series* (pp. 29–52). Springer International Publishing.

Bloom, B. S. (1968). *Learning for mastery.* UCLA-CSEIP Evaluation Comment. *1*(2), 1–12.

Bloom, B. S. (1984). *The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Brown, T. (2008). Design thinking: Thinking like a designer can transform the way you develop products, services, processes—and even strategy. *Harvard Business Review, 86*(6), 84–92.

Clements, D. H. (2007). Curriculum research: Toward a framework for research-based curricula. *Journal for research in mathematics education*, *38*(1), 35–70.

Dohring, D. C., Hendry, D. A., Gunderia, S., Hughes, D., Owen, V. E., Jacobs, D. E., Betts, A., & Salak, W. (2019). *Personalized mastery learning platforms, systems, media, and methods* (U.S. Patent No. 10490092). U.S. Patent and Trademark Office.

Dohring, D. C., Hendry, D. A., Gunderia, S., Hughes, D., Owen, V. E., Jacobs, D. E., Betts, A., & Salak, W. (2021). *System and method for dynamically editing online interactive elements architecture* (U.S. Patent No. 11151887). U.S. Patent and Trademark Office.

Dohring, D. C., Hendry, D. A., Gunderia, S., Hughes, D., Owen, V. E., Jacobs, D. E., Betts, A., & Salak, W. (2022). *Personalized mastery learning platforms, systems, media, and methods* (U.S. Patent No. 11380211). U.S. Patent and Trademark Office

Falmagne, J.-C., & Doignon, J.-P. (1999). *Learning spaces*. Springer.

The Gordon Commission on the Future of Assessment in Education. (2013). *To Assess, To Teach, To Learn: A Vision for the Future of Assessment, Technical Report Executive Summary.* ETS

Gunderia, S. (2024, January 18). The urgent call for educational innovation. *Fast Company*. https://www.fastcompany.com/91011298/the-urgent-call-for-educational-innovation

Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: Impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia, 12*(1), 40. https://doi.org/10.1186/s40468-022-00191-4

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. 16). Educational Testing Service. http://marces.org/EDMS623/Mislevy%20on%20ECD.pdf

(NAEP) National Center for Education Statistics. (n.d.). *NAEP mathematics: National achievement levels*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/mathematics/nation/achievement/?grade=4

(NAEP) National Center for Education Statistics. (n.d.). *NAEP reading: National achievement levels*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/reading/nation/achievement/?grade=4

(OECD) Organisation for Economic Co-operation and Development. (2023). *PISA 2022 results (Volume I & II) - Factsheets: United States*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Owen, V. E., & Baker, R. S. (2019). Learning analytics for games. In J. L. Plass, R. Meyer, & B. D. Homer (Eds.), *Handbook of Game-Based Learning* (pp. 513–535). MIT Press.

Owen, V. E., Roy, M.-H., Thai, K., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019). Detecting wheel-spinning and productive persistence in educational games. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 378–383). International Educational Data Mining Society.

Pape, B. (2018). *Learner variability is the rule, not the exception.* Digital Promise. https://digitalpromise.org/wp-content/uploads/2018/06/Learner-Variability-Is-The-Rule.pdf

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40(6), 601-618.

Rose, T. (2016). *The end of average: How we succeed in a world that values sameness*. HarperOne.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. MIT Press.

Thai, K. P., Bang, H. J., & Li, L. (2021). Accelerating Early Math Learning with Research-Based Personalized Learning Games: A Cluster Randomized Controlled Trial. *Journal of Research on Educational Effectiveness, 15*(1), 28–51.

Thai, K. P., Betts, A., & Gunderia, S. (2022). Personalized mastery-based learning ecosystem: A new paradigm for improving outcomes and defying expectations in early childhood. In A. Betts & K.P. Thai (Eds.), *Handbook of Research on Innovative Approaches to Early Childhood Education and School Readiness* (pp. 29-52). IGI Publishing.

Thai, K. P., Buchan, S., Kates, A., Blinder, E., Zeirath, C., & Betts, A. (2022). EdTech for "littles": Using a human-centered design approach to create a digital math readiness program for 2- to 4-year-olds. In R. K. Sawyer (Ed.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 2759–2792). Springer International Publishing.

United States Department of Education. (2025, January 29). U.S. Department of Education issues Statement on the Nation's Report Card. https://www.ed.gov/about/news/press-release/us-department-of-education-issues-statement-nations-report-card

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Handbook for Assessment in the Service of Learning Volume III Afterword

Eva L. Baker, Howard T. Everson, and Eric M. Tucker

What is the impact of this Volume? For decades, educators have aspired to transform assessment from a means of evaluating students' progress into a catalyst for learning—a vision championed by Edmund W. Gordon, who argues that the act of assessing should cultivate deeper understanding (Gordon Commission on the Future of Assessment in Education, 2013; Gordon, 2020). The collection of chapters in this volume reflects a shift from articulating the "why" to demonstrating the "how"; from aspiration to application. Building upon the theoretical and conceptual foundations set out in Volumes I and II, Volume III presents working examples of how to engineer and integrate assessment into learning across early childhood programs, K–12 classrooms, secondary science classrooms and art studios, and online platforms.

Taken together, these examples braid the design principles articulated in Volume I by Baker and colleagues—transparency; explicit focus and purpose; support for learner processes (motivation, attention, engagement, metacognition); modeling of growth over time; feedback-linked instruction; equity and attention to learner variation; and quality and validity (Baker, Everson, Tucker, & Gordon, 2025). The chapter authors demonstrate how educational assessments can be powerful learning experiences for students, while also providing instructionally useful information to teachers, students, and other stakeholders. These efforts intentionally embrace evidence-based design principles. Pellegrino and Everson (2025), for example, present a Next Generation Science Standards-aligned assessment approach that informs middle-school science instruction. Baker and Chung (2025), Buckley and Snow (2025), and DiCerbo (2025) show that game-based assessments yield rich evidence while keeping learners engaged.

By gathering evidence in instructional settings and providing useful and timely feedback, these examples show how assessment can be integrated with teaching and learning (Baker & Gordon, 2014).

As *evidentiary arguments*, these examples support warranted inferences about student competencies. Recognizing the role of feedback, they deliver timely, actionable insights to advance students' understanding and skills. As a *social practice*, the assessments are situated in authentic teaching and learning contexts. The exemplars show that when designed to be culturally responsive and identity-affirming—such as the AP Art and Design portfolio—assessments can become instruments of opportunity (Stone, Escoffery, Tabony, & Packer, 2025). Similarly, co-designing early childhood measures with communities can ensure assessments are culturally relevant, useful, and fair (Hanno, Mokyr Horner, Portilla, & Hsueh, 2025)

Conclusion: To Assess is to Teach and to Learn

Ultimately, this volume demonstrates that the maxim "to assess is to teach and to learn" translates into practical design principles (Baker et al., 2025). The path forward requires test developers, educators, and policymakers to consider the structural assets and cultural conditions for these innovations to thrive. To inform those discussions, this volume provides a collection of practical playbooks for architects of future assessment systems working to assemble the models and tools necessary to cultivate learning and drive meaningful improvement at both the classroom and system levels. The goal is not merely to create better tests. It is, as Gordon reminds us, to design educational assessments so that learners engage with them in ways that catalyze and cultivate learning and build enduring skills.

References

Baker, E. L., Everson, H. T., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Baker, E. L., & Gordon, E. W. (2014). *From the assessment of education to the assessment for education: Policy and futures. Teachers College Record, 116,* 1–24.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical Report.* Educational Testing Service.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In S. G. Sireci, E. M. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume II: Reconceptualizing Assessment to Improve Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W., & Bridglall, B. L. (Eds.). (2006). *Affirmative development: Cultivating academic ability* (Critical Issues in Contemporary American Education). Rowman & Littlefield.

Gordon, E. W., & Rajagopalan, K. (2016). New approaches to assessment that move in the right direction. In *The testing and learning revolution: The future of assessment in education* (pp. 107–146). Palgrave Macmillan.

Principles for Assessment Design and Use in the Service of Learning

This page outlines principles that guide the design and use of learning-focused assessments intended to support student learning. In the Handbook volumes, the principles were intended to assist chapter authors in considering these common elements in their contributions.

- **Principle 1:** Assessment **transparency** provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.
- **Principle 2:** Assessment **focus** is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.
- **Principle 3:** Assessment design supports learners' **processes**, such as motivation, attention, engagement, effort, and metacognition.
- **Principle 4:** Assessments model the structure of **expectations** and **desired learning** over time.
- **Principle 5: Feedback**, adaptation, and other relevant instruction should be linked to assessment experiences.
- **Principle 6:** Assessment **equity** requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.
- **Principle 7:** Assessment **quality and validity** should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

For those interested in the scientific or experiential bases of the principles, we refer you to the selected bibliography below. For each principle, the selected bibliography provides a set of references that highlight its theoretical and empirical underpinnings.

For more information, please refer to:

Baker, E. L., Everson, H. T., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Selected Bibliography

Assessment in the Service of Learning

Baker, E. L., & Gordon, E. W. (2014). From the assessment of education to the assessment for education: Policy and futures. *Teachers College Record, 116,* 1–24.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st-century learning.* Jossey-Bass.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical Report]. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Pellegrino, J. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116*(110313). https://doi.org/10.1177/016146811411601102

Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom activity systems to supportambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Principle 1: Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the I/ITSEC, 25,* 1811–1822.

Clancey, W. J., & Shortliffe, E. H. (Eds.). (1984). *Readings in medical artificial intelligence: The first decade.* Addison-Wesley. https://impact.dbmi.columbia.edu/~ehs7001/Clancey-Shortliffe-1984/Readings%20Book.htm

Gagné, R. M., & Briggs, L. J. (1974). *Principles of instructional design.* Holt, Rinehart & Winston.

Iseli, M. R., & Jha, R. (2016). Computational issues in modeling user behavior in serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 21–40). Routledge.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840–852. https://doi.org/10.1080/02602938.2013.875117

Moss, C. M., & Brookhart, S. M. (2012). *Learning targets: Helping students aim for understanding in today's lesson.* ASCD.

Principle 2: Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). Longman.

Andrade, H. L., Bennett, R. E., & Cizek, G. J. (Eds.). (2019). *Handbook of formative assessment in the disciplines* (1st ed.). Routledge. https://doi.org/10.4324/9781315166933

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of dynamic pedagogy: An integrative model of curriculum instruction and assessment for prospective and in-service teachers.* Routledge.

Chatterji, M. (2025). *User-centered assessment design: An integrated methodology for diverse populations.* Guilford Press.

Heritage, M. (2021). *Formative assessment: Making it happen in the classroom* (2nd ed.). Corwin.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education, 67*(3), 268–279. https://www.jstor.org/stable/2668195?origin=crossref

van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning: A systematic approach to four-component instructional design.* Routledge.

Principle 3: Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261–271.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066x.34.10.906

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in Cognitive Load Theory. *Educational Psychology Review, 31*(2), 339–359. https://doi.org/10.1007/s10648-019-09473-5

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Principle 4: Assessments Model the structure of expectations and desired learning over time.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives, 9*(2–3), 71–123.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, W. J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, L. A., Shulman, L., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

Gordon, E. G., & Bridgall B. L. (Eds.). (2006). *Affirmative development: Cultivating academic ability, critical issues in contemporary American education series*. Rowman & Littlefield Publishers, Inc.

Leonard, W. H., & Lowery, L. F. (1984). The effects of question types in textual reading upon retention of biology concepts. *Journal of Research in Science Teaching, 21*(4), 377–384. https://doi.org/10.1002/tea.3660210405

Phelps, R. P. (2012). The effects of testing on student achievement, 1910–2010. *International Journal of Testing, 12,* 21–43.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Principle 5: Feedback, adaptation, and other relevant instruction should be linked to assessment experiences.

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement* (1st ed.). Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97. https://doi.org/10.3102/00346543058001079

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179–189.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10,* Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Principle 6: Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.

Armour-Thomas, E., McCallister, C., Boykin, A. W., & Gordon, E. W. (Eds.). (2019). *Human variance and assessment for learning.* Third World Press.

Bennett, R. E. (2023). *T*oward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). Macmillan.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 360–372.

Herman, J. L., Bailey, A. L., & Martinez, J. F. (2023). Introduction to the special issue: Fairness in educational assessment and the next edition of the standards. *Educational Assessment, 28*(2), 65–67. https://doi.org/10.1080/10627197.2023.2215979

Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning.* Routledge.

Oakes, J. (1986). Keeping track, part 1: The policy and practice of curriculum inequality. *Phi Delta Kappan, 68*(1), 12–17. https://www.jstor.org/stable/20403250

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator, 45*(3), 28–37, 48.

Sireci, S. G., & Randall, J. (2021). *Evolving notions of fairness in testing in the United States in the history of educational measurement.* Routledge.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13. https://doi.org/10.3102/0013189x032002003

Principle 7: Quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Linn, R. L. (2010). Validity. In B. McGaw, P. L. Peterson, & E. L. Baker (Eds.), *International encyclopedia of education* (3rd ed., Vol. 4, pp. 181–185). Elsevier.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Mislevy, R. J., Oliveri, M. E., Slomp, D., Crop Eared Wolf, A., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 199–241). Routledge/Taylor & Francis.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–67.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99–104.

Series Contributors

Sergio Araneda, University of Massachusetts Amherst

Eleanor Armour-Thomas, Queens College, City University of New York (Emeritus)

Aneesha Badrinarayan, Education First

Eva L. Baker, University of California, Los Angeles, Center for Research on Evaluation, Standards & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Hee Jin Bang, Age of Learning

Héfer Bembenutty, Queens College, City University of New York

Randy E. Bennett, ETS, Research Institute

Anastasia Betts, Learnology Labs

Mary K. Boudreaux, Southern Connecticut State University

Susan M. Brookhart, Duquesne University

Carol Bonilla Bowman, Ramapo College of New Jersey

Jack Buckley, Roblox

Jill Burstein, Duolingo, Inc.

Pamela Cantor, The Human Potential L.A.B.

Jennifer Charlot, RevX

Gregory K. W. K. Chung, University of California, Los Angeles, Center for Research on Evaluation, Standards & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Paul Cobb, Vanderbilt University

Kimberly Cockrell, The Achievement Network, Ltd.

Kelly Corrado, PBS KIDS

Danielle Crabtree, University of Massachusetts Amherst

Linda Darling-Hammond, Learning Policy Institute

Jacqueline Darvin, Queens College, City University of New York

Girlie C. Delacruz, Northeastern University

Clarissa Deverel-Rico, BSCS Science Learning

Kristen DiCerbo, Khan Academy

Ravit Dotan, TechBetter LLC

Kerrie A. Douglas, Purdue University

Kadriye Ercikan, Educational Testing Service

David S. Escoffery, Educational Testing Service

Series Contributors (continued)

Carla M. Evans, National Center for the Improvement of Educational Assessment

Howard T. Everson, Graduate Center, City University of New York

Cosimo Felline, PBS KIDS

Kate Felsen, The Human Potential L.A.B.

Tianying Feng, University of California, Los Angeles, Center for Research on Evaluation, Standards & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Natalie Foster, Organisation for Economic Co-operation and Development (OECD)

James Paul Gee, Arizona State University (Emeritus)

Sheryl L. Gómez, The Study Group

Edmund W. Gordon, Teachers College, Columbia University (Emeritus); Yale University (Emeritus)

Sunil Gunderia, Age of Learning

Laura S. Hamilton, National Center for the Improvement of Educational Assessment

Emily C. Hanno, MDRC

John Hattie, University of Melbourne (Emeritus)

Norris M. Haynes, Southern Connecticut State University

JoAnn Hsueh, MDRC

Kristen Huff, Curriculum Associates

Diana Hughes, Relay Graduate School of Education

Gerunda B. Hughes, Howard University (Emeritus)

Neal Kingston, University of Kansas

Geoffrey T. LaFlair, Duolingo, Inc.

Carol D. Lee, Northwestern University (Emeritus)

Paul G. LeMahieu, Carnegie Foundation for the Advancement of Teaching; University of Hawai'i, Mānoa

Richard M. Lerner, Tufts University

Lei Liu, Educational Testing Service

Ou Lydia Liu, Educational Testing Service

Silvia Lovato, PBS KIDS

Temple S. Lovelace, Assessment for Good, Advanced Education Research and Development Fund (AERDF)

Susan Lyons, Lyons Assessment Consulting

Scott F. Marion, National Center for the Improvement of Educational Assessment

Kimberly McIntee, University of Massachusetts Amherst

Maxine McKinney de Royston, Erikson Institute

Elizabeth Mokyr Horner, Gates Foundation

Orrin T. Murray, The Wallis Research Group

Na'ilah Suad Nasir, Spencer Foundation

Michelle Odemwingie, The Achievement Network, Ltd.

Maria Elena Oliveri, Purdue University

Saskia Op den Bosch, RevX

V. Elizabeth Owen, Age of Learning

Trevor Packer, College Board

Roy Pea, Stanford University

James W. Pellegrino, University of Illinois Chicago

Mario Piacentini, Organisation for Economic Co-operation and Development (OECD)

Mya Poe, Northeastern University

Ximena A. Portilla, MDRC

Elizabeth J. K. H. Redman, University of California, Los Angeles, Center for Research on Evaluation, Standards & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Jeremy D. Roberts, PBS KIDS

Mary-Celeste Schreuder, The Achievement Network, Ltd.

David Sherer, Carnegie Foundation for the Advancement of Teaching

Stephen G. Sireci, University of Massachusetts Amherst, Center for Educational Assessment

Erica Snow, Roblox

Rebecca A. Stone-Danahy, College Board

Rebecca Sutherland, Reading Reimagined, Advanced Education Research and Development Fund (AERDF)

Natalya Tabony, College Board

Carrie Townley-Flores, Rapid Online Assessment of Reading (ROAR), Stanford University

Eric M. Tucker, The Study Group

Alina A. von Davier, Duolingo, Inc.

Kevin Yancey, Duolingo, Inc.

Jessica W. Younger, PBS KIDS

Constance Yowell, Northeastern University

Biographical Statements

Sergio Araneda, Ph.D., is a research scientist specializing in educational measurement, psychometrics, and test security. He earned his doctorate in Research, Educational Measurement, and Psychometrics from the University of Massachusetts Amherst, following completion of his undergraduate studies in Mathematical Civil Engineering at the Universidad de Chile. Dr. Araneda currently works at Caveon, where he investigates how large language models can be integrated with test security innovations such as SmartItems™, contributing to research, publications, and conference presentations. He previously served as an associate psychometrician at the College Board, focusing on item parameter drift and automated essay scoring for the SAT, and as a research assistant at DEMRE, Universidad de Chile, evaluating policies in university admissions. His earlier professional experience also includes roles in finance as a quantitative analyst and consultant, providing him with a strong technical and analytical background. His academic and professional contributions span peer-reviewed publications, white papers, newspaper columns, and numerous presentations at international conferences, including NCME, ITC, and ATP. He also serves as Vice-Coordinator of FEVED, a professional forum advocating for best practices in educational assessment in Chile.

Eleanor Armour-Thomas, Ed.D., is Professor Emerita at Queens College, CUNY, where she served in the Department of Secondary Education from 1987 to 2024, including 22 years (2000–2022) as Department Chair. She specialized in Educational Psychology, teaching pre-service and in-service teachers, and served as Principal Investigator and Co-Principal Investigator for programs aimed at enhancing mathematics teacher preparation and professional development in science education. Her books, journal articles, oral addresses, and reports focus on teacher and student cognition, metacognition, learning, and assessment. Additionally, she has evaluated educational programs designed to improve learning and academic achievement for students from low socio-economic backgrounds and has consulted on teaching, learning, and assessment in K-16 education.

Aneesha Badrinarayan is a Principal Consultant at Education First, where she partners with state and district leaders, assessment developers, and policymakers to design coherent systems of teaching, learning, and assessment. She brings decades of expertise in assessment design, STEM education, policy, and product development, helping organizations and leaders create and implement instructionally relevant assessment systems. At Education First, Aneesha leads projects on innovative assessment and accountability design, equitable assessment, strategic planning, and artificial intelligence. Previously, Aneesha directed assessment work at the Learning Policy Institute, leading innovations across 15 states, shaping the 2028 NAEP Science Framework, and guiding federal policy on learning-first assessments. A behavioral neuroscientist by training, she holds degrees from Cornell University and the University of Michigan.

Eva L. Baker is a Distinguished Professor at UCLA and founding Director of the Center for Research on Evaluation, Standards and Student Testing, (CRESST). She is widely published in the areas of learning-based assessments, technology, and policy. She served as Chair of the Board on Testing and Assessment, National Research Council, and Co-Chair of the 1999 Standards for Educational and Psychological Testing. Baker served as president of the World Education Research Association (WERA) and was president of the American Educational Research Association (AERA). A member of the National Academy of Education, she received AERA's Robert L. Linn Lecture and the E. F. Lindquist Award.

Dr. Hee Jin Bang, Vice President of Efficacy Research & Evaluation at Age of Learning, Inc., leads research initiatives evaluating the effectiveness of educational technology products. In her current role, she oversees research studies examining the impact of adaptive learning technologies on student achievement across diverse populations and educational settings. Her recent publications offer compelling evidence for the effectiveness of digital learning platforms, demonstrating significant learning gains in language acquisition, early mathematics, and reading skills. Currently, as co-principal investigator on a $3.5 million Institute of Education Sciences-funded study, she continues to shape more effective educational technology solutions by investigating how personalized game-based learning supports teaching and learning in classrooms. Prior to joining Age of Learning, she held research leadership positions at Classroom, Inc., Amplify Education, and National Writing Project, where she evaluated digital curricula, assessments, and teacher professional development programs. She holds a Ph.D. from NYU in Teaching & Learning, an M.Ed. in Human Development and Psychology from Harvard University, and a B.A. (Honors) in Linguistics and French from Oxford University.

Héfer Bembenutty, Ph.D., is dedicated to advancing the field of educational psychology through his role as a professor at Queens College, The City University of New York. His academic journey led him to earn a Ph.D. in educational psychology from the same institution. Dr. Bembenutty's research focuses on the self-regulation of learning among high school and college students, as well as teachers. He explores various aspects such as assessment, homework self-regulation, self-efficacy beliefs, culturally self-regulated pedagogy, and academic delay of gratification. His teaching portfolio includes undergraduate and graduate courses on educational psychology, cognition, instruction and technology, human development and learning, assessment and measurement, and classroom management. Additionally, he investigates the impact of demographic factors like gender and ethnicity on students' ability to prioritize long-term goals over immediate rewards. He is an accomplished author and editor, contributing to several books and peer-reviewed journals. His work integrates contemporary theories with practical applications to enhance self-regulated learning in educational environments.

Randy E. Bennett holds the Norman O. Frederiksen Chair in Assessment Innovation in the ETS Research Institute. His recent work centers on personalized assessments and, relatedly, assessments that are "born socioculturally responsive." From 1999–2005 he directed the National Assessment of Educational Progress (NAEP) Technology-Based Assessment project, which included the first administration of computer-based performance assessments to nationally representative samples of U.S. school students and the first use of logfile data in such samples to measure problem-solving processes. From 2007–2016, he directed the CBAL research initiative (Cognitively Based Assessment of, for, and as Learning), which created theory-based summative and formative assessment to model good teaching and learning practice. He is a past president of the International Association for Educational Assessment and of the National Council on Measurement in Education (NCME). He is a fellow of the American Educational Research Association (AERA) and an elected member of the National Academy of Education, as well as recipient of the NCME Bradley Hanson Contributions to Educational Measurement Award, the Teachers College Columbia University Distinguished Alumni Award, the AERA E. F. Lindquist Award, and the AERA Cognition and Assessment SIG Award for Outstanding Contribution to Research in Cognition and Assessment.

Dr. Anastasia Betts is a leading expert in education and learning sciences innovation. As Executive Director of Learnology Labs, a collaborative think tank, she leads cutting-edge research on AI-enabled learning systems with a focus on transforming early childhood. Dr. Betts previously led the curriculum research, design, and production of digital learning products for early learning at Age of Learning, where her pioneering work in adaptive learning systems resulted in her inclusion on three U.S. patents. Currently, Dr. Betts spearheads the development of PAL (Personal Assistant for Learning), an AI-driven system that exemplifies distributed cognition principles to empower parents and teachers in supporting early math development. Dr. Betts holds a Ph.D. in Curriculum, Instruction, & the Science of Learning from the University at Buffalo, SUNY. Her research and publications focus on leveraging learning sciences and AI to create more equitable, personalized educational experiences. She is editor of the *Handbook of Research for Innovative Approaches to Early Childhood Education and Kindergarten Readiness* and has authored numerous papers on adaptive learning and human-AI partnerships in education. Dr. Betts was selected as a Harvard Women in Educational Leadership Fellow and was twice nominated for the American Educational Research Association (AERA) Karen King Future Leader Award.

Dr. Mary K. Boudreaux is an Associate Professor and Coordinator of the Doctoral Program in Educational Leadership & Policy Studies at Southern Connecticut State University. With a distinguished career spanning K–12 and higher education, she has served as a curriculum director, educational specialist, consultant, and university faculty member. Dr. Boudreaux specializes in improving school culture and climate, enhancing leadership practices, and promoting equity-focused practices and assessment strategies. As an educator and scholar, Dr. Boudreaux has designed and taught graduate and doctoral courses in organizational leadership, research methods, curriculum development, assessment, and change leadership. Her work prepares aspiring and practicing educational leaders to address systemic challenges through data-driven decision-making and evidence-based assessment practices. A prolific researcher, she has published numerous peer-reviewed articles, book chapters, and conference presentations on multicultural awareness and leadership, as well as fostering inclusive and equitable learning environments. Dr. Boudreaux's commitment to continuous improvement in education is reflected in her leadership roles as Co-Chair of the University Standards and Assessment Review Committee and a member of the University Graduate Council. These positions allow her to shape institutional assessment practices, ensuring academic programs achieve and maintain high-quality performance standards. Holding doctoral degrees in Educational Leadership and Innovation and Curriculum & Instruction, alongside certifications in higher education leadership, instructional design, and academic advising, Dr. Boudreaux remains dedicated to enhancing educational excellence and shaping future generations of scholars and practitioners.

Susan M. Brookhart, Ph.D., is Professor Emerita in the School of Education at Duquesne University and an independent educational consultant. She was the 2007-2009 Editor of Educational Measurement: Issues and Practice and is currently an Associate Editor of Applied Measurement in Education. She is the author or co-author of over 100 articles, chapters, and books on classroom assessment, teacher professional development, and evaluation. She was named the 2014 Jason Millman Scholar by the Consortium for Research on Educational Assessment and Teaching Effectiveness (CREATE) and was the recipient of the 2015 Samuel J. Messick Memorial Lecture Award from ETS/TOEFL. Dr. Brookhart's research interests include the role of both formative and summative classroom assessment in student motivation and achievement, the connection between classroom assessment and large-scale assessment, and grading. Dr. Brookhart received her Ph.D. in Educational Research and Evaluation from The Ohio State University, after teaching in both elementary and middle schools.

Dr. Carol Bonilla Bowman is an Associate Professor of Education at Ramapo College of New Jersey, where she also serves as a program director. Her research and publications focus on portfolios as both assessment and learning tools. Her recent work focuses on contemplative education. She holds a doctoral degree in applied linguistics and bilingual education from Teachers College, Columbia.

Dr. Sean P. "Jack" Buckley is Vice President of People at Roblox, where he oversees several teams including People (HR) and People Science and Analytics. He was previously President and Chief Scientist at Imbellus, Senior Vice President at the American Institutes for Research (AIR), and Senior Vice President of Research at The College Board. He also served as Commissioner of the U.S. Department of Education's National Center for Education Statistics (NCES) and as an Associate Professor at New York University, and an Assistant Professor at Boston College. He began his career as a surface warfare officer and nuclear reactor engineer in the U.S. Navy and has also worked in intelligence analysis. He holds an M.A. and Ph.D. in Political Science from Stony Brook University and an A.B. in Government from Harvard University.

Jill Burstein is Principal Assessment Scientist at Duolingo, leading validity and efficacy research for the Duolingo English Test – Duolingo's English language proficiency test. Her career has been motivated by social impact, working on AI-driven, education technology to enhance equity and access for learners and test-takers. Her research lies at the intersection of artificial intelligence and natural language processing, educational measurement, equity in education, learning analytics, and linguistics. Dr. Burstein pioneered the first automated writing evaluation system used in large-scale, high-stakes assessment, as well as early commercial online writing instruction tools. She holds numerous patents for this work, and has published extensively in the field of AI in education, including topics in automated writing evaluation, digital assessment, responsible AI, and writing analytics. Her recent work focuses on responsible AI for digital assessment, and wrote the Duolingo English Test Responsible AI Standards, the first standards for an assessment program. Additionally, she is a co-founder of SIG EDU, an ACL Special Interest Group on Building Educational Applications. Dr. Burstein holds a Ph.D. in Linguistics from the Graduate Center, City University of New York.

Pamela Cantor, M.D., is a child and adolescent psychiatrist and the Founder and CEO of The Human Potential L.A.B., whose mission is to leverage scientific knowledge and technologies to transform what people understand and what institutions do to unlock human potential in each and every individual. Dr. Cantor is an author of Whole-Child Development, Learning and Thriving: A Dynamic Systems Approach (Cambridge University Press) and The Science of Learning and Development (Routledge). She founded the nonprofit organization Turnaround for Children (now the Center for Whole-Child Education at Arizona State University), is a Governing Partner of the Science of Learning and Development Alliance, and a strategic science advisor to the Carnegie Foundation for the Advancement of Teaching, the American Association of School Superintendents, and Learning Heroes. Dr. Cantor received an M.D. from Cornell University, a B.A. from Sarah Lawrence College, served as an Assistant Clinical Professor of Child Psychiatry at Yale School of Medicine, and was a Visiting Scholar at the Harvard Graduate School of Education.

Dr. Jennifer Charlot is co-founder of RevX, where she serves as Head of Programming. She leads the implementation of RevX's assessment system, ensuring data collection is integrated into daily instruction and shaping our systems for using real-time insights to refine teaching practice. As Managing Partner at Transcend, she directed early-stage school design projects, spreading science-driven innovation nationwide. A serial entrepreneur, Dr. Charlot spearheaded career and technical education programs for disconnected youth in NYC and served as Director of Implementation at Character Lab, translating research into practical classroom strategies. She holds a Doctorate in Education Leadership from Harvard's Graduate School of Education, a Master of Science in Social Administration from Columbia University, and a Bachelor of Arts from Boston College. Dr. Charlot is dedicated to reimagining educational systems through innovative design, actionable strategies, and data-driven practice—empowering young people to emerge as changemakers in their communities and beyond.

Gregory K. W. K. Chung, Ph.D. is the Associate Director for Technology and Research Innovation. Dr. Chung has extensive experience with the use of technology for learning and assessment. He has led projects related to game-based learning or game-based assessments involving pre-school students to adults in formal and informal settings with a focus on STEM topics (e.g., math, physics, engineering, programming) as well as social-emotional learning. His research involves small-scale exploratory studies to multi-district, multi-state RCT. He has conducted instructional technology R&D for IES, NSF, Office of Naval Research, PBS KIDS, Bill and Melinda Gates Foundation, Caplan Foundation for Early Childhood, and numerous other foundations and commercial entities.

Paul Cobb is Professor Emeritus at Vanderbilt University. His work focuses on improving the quality of mathematics teaching and student learning on a large scale. He is currently involved in a project that is developing practical measures of key aspects of high quality mathematics and investigating their use as levers for and measures of instructional improvement. He received Hans Freudenthal Medal for cumulative research program over the prior ten years from the International Commission on Mathematics Instruction (ICMI) in 2005, and the Silver Scribner Award from American Educational Research Association in 2010 for research over the past ten years that contributes to our understanding of learning and instruction.

Kimberly Cockrell is an experienced educator, administrator, and leader committed to instructional excellence, leadership development, and equity in education. With over two decades of experience in school leadership, professional learning, and strategic partnerships, she has worked to transform assessment and instructional practices to better support educators and students. At Achievement Network (ANet), Kimberly directs communications and stakeholder engagement, shaping public discourse around instructional coherence, data-driven decision-making, and student success. Kimberly's career spans charter, public, and independent schools, where she has designed professional development programs, led data-driven instructional strategies, and championed equitable learning environments. A lifelong learner and consultant, she continues to support educators in strengthening school leadership, assessment literacy, and instructional coherence.

Kelly Corrado is the Director of Game Tooling and Analytics Products for PBS KIDS. Corrado is committed to leveraging technology to enrich early childhood education through the delivery of high-impact products and experiences at scale for children aged 2-8 and the grownups who support them in school and in life. Corrado is a results-driven product leader with success leading cross-functional teams, optimizing digital ecosystems, and driving strategic initiatives that enhance accessibility, performance, and engagement. With a focus on game development and analytics platforms, Corrado influences business growth and user experience through data insights and innovation.

Danielle Crabtree, M.Ed., is a doctoral student in the Research, Educational Measurement, and Psychometrics program at the University of Massachusetts Amherst. She holds dual master's degrees in Educational Administration and Secondary Education, and bachelor's degrees in Mathematics and Biochemistry & Molecular Biology, giving her a strong interdisciplinary foundation. Her research examines educational equity, teacher professional learning, and technology-enhanced instruction. She focuses on developing new methods to capture complex, hidden aspects of teaching and learning, broadening how assessment can inform both research and practice. As a Graduate Research Assistant, Danielle has contributed to WearableLearning, a game-based platform integrating embodied learning and computational thinking in mathematics led by Professor Ivon Arroyo, and EMPOWER, a research-practice partnership exploring the development of teacher educators' critical consciousness in science classrooms led by Associate Professor Enrique Suárez. She has co-authored multiple peer-reviewed conference proceedings, including a 2024 paper nominated for Best Design Paper at the International Conference of the Learning Sciences. An experienced educator and administrator, Danielle has served as a classroom teacher, assistant principal, practicum supervisor, and university instructor. She holds licensure as both a secondary teacher and PreK–12 principal. Passionate about advancing educational equity and innovation, she works to bridge research and practice to strengthen teacher development and improve outcomes for both teachers and students.

Linda Darling-Hammond is the Charles E. Ducommun Professor of Education, Emeritus, at Stanford University and founding president of the Learning Policy Institute, where she leads research and policy initiatives focused on educational equity, teacher quality, and effective school reform. A nationally renowned scholar, she has authored more than 30 books and hundreds of publications on teaching, learning, and education policy. Darling-Hammond's career has centered on advancing evidence-based policies that improve access to high-quality learning opportunities for all students. She served as chair of the California State Board of Education from 2019 to 2023, where she guided the state's efforts to strengthen curriculum, assessments, and teacher preparation. Earlier, she directed the Stanford Center for Opportunity Policy in Education and the National Commission on Teaching and America's Future, influencing reforms in teacher development and accountability systems across the U.S. Recognized as one of the most influential voices in education, she has advised federal and state leaders on issues ranging from school funding to equitable assessment design. Darling-Hammond continues to champion the creation of schools that support deep learning, social-emotional growth, and equitable outcomes for every child.

Jacqueline Darvin, Ph.D., is a Program Director and Professor of Literacy Education at Queens College of the City University of New York (CUNY). In addition to a BA in Psychology and doctorate in Literacy Studies, she has master's degrees in educational leadership and secondary education and credentials as a New York State School District Leader. Before becoming a professor at Queens College, Dr. Darvin taught middle and high school Title One reading, Special Education, and English for twelve years. In 2015, she published a book with Teachers College Press titled Teaching the Tough Issues: Problem-Solving from Multiple Perspectives in Middle and High School Humanities Classes. She was the recipient of the Long Island Educator of the Month Award, featured in a cover story of New York Teacher, the official publication of the New York State United Teachers' Union, and a recipient of the Queens College Presidential Award for Innovative Teaching. She is a workshop provider for Nassau and Easter Suffolk BOCES and provides consulting and professional development to schools and teachers throughout the New York metropolitan area. Her presentations include local, regional, national and international conferences on topics related to literacy teaching and learning.

Girlie C. Delacruz is Associate Vice Chancellor for Teaching and Learning at Northeastern University, where she oversees experiential learning programs in undergraduate research, service learning, and community and civic engagement, as well as student support through fellowships advising and peer tutoring. With over two decades of experience spanning research and applied practice, she has led initiatives to expand equitable access to education, including as Chief Learning Officer for LRNG at Southern New Hampshire University and as a researcher at UCLA developing technology-enhanced assessments for military and educational contexts. Her scholarship and leadership have been recognized through awards such as Northeastern's 2025 Staff Excellence Award for Mentorship and the APA Military Psychology Research Award, as well as fellowships from the MacArthur Foundation and ETS. She also serves on national grant review panels and has published widely on learning, assessment design, and the role of technology in advancing equity.

Clarissa Deverel-Rico, Ph.D., is a postdoctoral researcher at BSCS Science Learning. A former middle-school science teacher, Clarissa transitioned into a career driven by creating better science learning experiences for students. She studies innovative approaches for how classroom assessment can support a vision of science education that prioritizes epistemic justice, care, and student experience. Current research aims include studying the extent to which currently available classroom assessments support equitable opportunities to learn, developing assessments for broad use in high school biology, investigating the efficacy of locally-adapted high-quality curricular materials, and partnering with teachers around creating spaces to learn directly from students and families for how classroom assessment can be spaces that sustain students' interests and identities.

Dr. Kristen DiCerbo is the Chief Learning Officer at Khan Academy, a nonprofit dedicated to providing a free world class education to anyone, anywhere. In this role, she is responsible for the research-based teaching and learning strategy for Khan Academy's offerings. She leads the content, assessment, design, product management, and community support teams. Time magazine named her one of the top 100 people influencing the future of AI in 2024. Dr. DiCerbo's work has consistently been focused on embedding what we know from education research about how people learn into digital learning experiences. Prior to her role at Khan Academy, she was Vice-President of Learning Research and Design at Pearson, served as a research scientist supporting the Cisco Networking Academies, and worked as a school psychologist in an Arizona school district. Kristen received her Bachelor's degree from Hamilton College and Master's degree and Ph.D. in Educational Psychology at Arizona State University.

Ravit Dotan, Ph.D., is a renowned tech ethicist specializing in artificial intelligence (AI) and data technologies. She aids tech companies, investors, and procurement teams in developing and implementing responsible AI strategies, conducts research on these topics and creates resources. Dr. Dotan was recognized as one of the 100 Brilliant Women in AI Ethics for 2023 and has received accolades such as the 2022 "Distinguished Paper" Award from the FAccT conference. Her views are frequently featured in prominent publications like the *New York Times, The Financial Times,* AP News, and TechCrunch. Dr. Dotan holds a Ph.D. in Philosophy from UC Berkeley and has extensive experience in AI ethics research, teaching, and advocacy for diversity and inclusion in academia. You can find Dr. Dotan's resources on her AI Ethics Treasure Chest and LinkedIn page.

Kerrie A. Douglas, Ph.D., is an Associate Professor of Engineering Education at Purdue University and Co-Director of SCALE, a large Department of Defense funded workforce development project in secure microelectronics. In that role, she leads the education and workforce development across 33 universities in the U.S. She is passionate about modernizing engineering education and preparing learners for their professional work. Her research is focused on improving methods of evaluation and assessment in engineering learning contexts. She works on assessment problems in engineering education, such as considerations for fairness, how to assess complex engineering competencies, and aligning assessment to emerging workforce needs. She has been Primary Investigator or Co-PI on more than $100 million of external research awards. In 2020, she received an NSF RAPID award to study engineering instructional decisions and how students were supported during the time of emergency remote instruction due to the COVID-19 pandemic. In 2021, she received the NSF CAREER award to study improving the fairness of assessment in engineering classrooms. She has published over 100 peer-reviewed journal and conference papers.

Dr. Kadriye Ercikan is the Senior Vice President of Global Research at the Educational Testing Service (ETS), President and CEO of ETS Canada Inc., and Professor Emerita at the University of British Columbia. In these leadership roles, she directs foundational and applied research. Her research focuses on validity and fairness issues and sociocultural context of assessment. Her recent research includes validity and fairness issues in innovative digital assessments, including using response process data, AI applications, and adaptivity. Ercikan is the President and a Fellow of the International Academy of Education (IAE), President of the International Test Commission (ITC), and President-Elect of the National Council on Measurement in Education (NCME). Her research has resulted in six books, four special issues of refereed journals and over 150 publications. She was awarded the AERA Division D Significant Contributions to Educational Measurement and Research Methodology recognition for another co-edited volume, Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization, and received an Early Career Award from the University of British Columbia. Ercikan is currently serving as the NCME Book Series Editor (2021-2026).

David S. Escoffery is a Director in the Graduate and Professional Education area at Educational Testing Service. He joined ETS in 2006 after teaching theatre history at the university level for five years. His academic areas of specialization include theatre history and literature, English language and literature, pedagogical theory, and cultural studies. He applies his experience to the development of examinations that measure knowledge of critical thinking, writing, and analytical reasoning. In addition to AP Art and Design, he has worked on a wide variety of assessment programs, including GRE, Praxis, and SAT. He has published numerous articles in journals such as Applied Measurement in Education and served as the editor for the 2006 McFarland collection How Real Is Reality TV? He earned his Ph.D. and M.A. in theatre history, literature, and criticism from the University of Pittsburgh, and his A.B. in English from Princeton University.

Carla M. Evans is a Senior Associate at the National Center for the Improvement of Educational Assessment, where she leads efforts to develop and implement balanced assessment and accountability systems for states, bridging the classroom and policymaking levels. Carla's work spans systemwide assessment reviews, assessment literacy initiatives, performance-based assessment design, and aligning accountability systems with educational values. Her research emphasis lies in culturally responsive assessment, competency-based education, AI in classroom assessment, and instructionally useful assessment.

Howard T. Everson is a Professor of Educational Psychology (by courtesy) at the Graduate School, City University of New York. He is the former Director of the Center for Advanced Study in Education at the Graduate School, City University of New York. His research and scholarly interests focus on the intersection of cognition, technology and assessment. He has published widely and has contributed to developments in educational psychology, psychometrics, quantitative methods, and program evaluation. Professor Everson's measurement expertise is in the areas of evidence-centered design, item response theory, differential item functioning, learning analytics and cognitive diagnostic measurement models. Dr. Everson also served as the Executive Director of the NAEP Educational Statistics Services Institute at the American Institutes for Research, and was the Vice President and Chief Research Scientist at the College Board. Dr. Everson is a Psychometric Fellow at the Educational Testing Service, and an elected Fellow of both the American Educational Research Association and the American Psychological Association, and a charter member of the Association for Psychological Science. Dr. Everson is the former editor of the National Council of Measurement in Education's journal, Educational Measurement: Issues and Practice.

Cosimo Felline, Ph.D., is the Director of Data Science and Analytics at PBS KIDS. With a background in theoretical nuclear physics, he earned his doctorate before transitioning from academia to the tech industry. Beginning his career as a web developer, software engineer, and manager, Felline developed a strong foundation in software development and web technologies. More recently, he has shifted his focus to data science and engineering, where he applies his expertise to building scalable data solutions. Passionate about data literacy and democratization, he is committed to breaking down barriers to data access and enabling actionable insights. He enjoys playing the piano, watching horror movies, and petting his dogs.

Kate Felsen is the Chief Communications Officer of The Human Potential L.A.B. and President of Up Up Communications LLC, with clients focused on transforming education and supporting healthy youth development. Kate had a distinguished career at ABC News. As Foreign Editor for the flagship evening news broadcast, she covered breaking and feature stories around the globe, winning 11 Emmy Awards. Kate earned an M.A. in American foreign policy and international economics from Johns Hopkins and a B.A., *magna cum laude* in history and literature from Harvard. She garnered first-team All-American and Ivy League "Player of the Year" honors in lacrosse, captained the field hockey team and enjoys coaching a club lacrosse team for middle-school girls in New York City. She serves as Chair of the Board of USA Climbing and Feed the Frontlines NYC.

Tianying Feng is a Ph.D. candidate in the Education – Advanced Quantitative Methods program at UCLA and a research assistant at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), SEIS Building, Los Angeles, CA 90095-1522; tfeng0315@ucla.edu. Her primary research interests include technology-based measurement and learning, psychometrics, process modeling, and statistical computing.

Natalie Foster is an Analyst in the Programme for International Student Assessment (PISA) at the Organisation for Economic Co-operation and Development (OECD). Her work mainly focuses on the design and development of innovative assessments of 21st century competences included in each PISA cycle, working closely with measurement and test development experts, as well as various other PISA research and development projects. She is the lead author of the PISA 2022 Creative Thinking and PISA 2025 Learning in the Digital World assessment frameworks, co-editor of the publication Innovating Assessments to Measure and Support Complex Skills, and the lead author of the PISA 2022 Results (Volume III): Creative Minds, Creative Schools report. She has also worked in the OECD Centre for Educational Research and Innovation on the Smart Data and Digital Technologies in Education project, where she contributed to the OECD Digital Education Outlook 2023. Before joining PISA, she worked at the OECD Development Centre and European Commission.

James Paul Gee is a Regents Professor Emeritus at Arizona State University. He was, in his career, a professor at six universities. He is an elected member of the National Academy of Education. He received his Ph.D. in linguistics in 1975 from Stanford University and initially worked on syntactic theory and the philosophy of language, later becoming interested in a variety of other areas, including psycholinguistics, discourse analysis, sociolinguistics, literacy studies, learning theory, and video games. His books include Sociolinguistics and Literacies; The Social Mind; An Introduction to Discourse Analysis; Situated Language and Literacies; What Video Games Have to Teach Us About Literacy and Learning; The Anti-Education Era; and What is a Human? His current work is about the paradox that while we say "humans learn from experience" and experience is composed of sensory interactions with the world, we hear precious little about sensation in educational research.

Sheryl L. Gómez, serves as the Chief Financial and Operating Officer for the Study Group, where she leads strategy, finance, and operations to advance equity, innovation, and impact in education. She is a results-driven finance and operations executive across the public, private, and social sectors. She has served as the CFO for Brooklyn Laboratory Charter Schools, CFO and COO of Friends of Brooklyn LAB, CFO and COO of Equity By Design, a Financial Manager at Charter School Business Management, and a Financial Manager at FOREsight Financial Services for Good. Her experience includes managing clients' accounts, maintaining accurate records of financial transactions, financial reports, monthly close reviews, financial audits, and year-end processes. She has expertise in organizational growth, resource development, financial strategy, and public-private partnerships. She has managed multimillion-dollar budgets, secured over $150M in facilities financing, and overseen grants from major funders.

Edmund W. Gordon is the John M. Musser Professor of Psychology, Emeritus at Yale University; Richard March Hoe Professor, Emeritus of Psychology and Education, at Teachers College, Columbia University; Director Emeritus of the Edmund W. Gordon Institute for Advanced Study, at Teachers College, Columbia University; and Honorary President of the American Educational Research Association. Gordon's distinguished career spans professional practice and scholarly life as a minister, clinical and counseling psychologist, research scientist, author, editor, and professor. He earned his B.S. in Zoology and B.D. at Howard University, an M.A. in Social Psychology from American University, and an Ed.D. in Child Development and Guidance from Teachers College, Columbia University. He received the AERA Relating Research to Practice Award (2010), the John Hope Franklin Award (2011), and the Harold W. McGraw, Jr. Prize in Education (2024). He is widely recognized for his work on the Head Start program, the achievement gap, supplementary education, the affirmative development of academic ability, and Assessment in the Service of Learning. Author of more than 400 articles and 25 books, Gordon has been named one of America's most prolific and thoughtful scholars. He was married to Susan Gitt Gordon for 75 years and together had four children.

Sunil Gunderia, is Chief Innovation Officer at Age of Learning, the company behind ABCmouse, an early learning program trusted by the parents of 50 million children. He co-invented the AI-based personalized mastery learning system powering My Math Academy and My Reading Academy, game-based programs whose effectiveness has been validated by 28 ESSA-aligned studies. Research finds over 90 percent of teachers want these programs for their impact on learning and on students' confidence and interest in reading and math. Sunil is Vice Chair of the EdSAFE AI Industry Council and Advisor to National AI Literacy Day and the Center for Outcome-Based Contracting. He also serves on the boards of InnovateEDU and the Children's Institute, which provides Head Start and mental health services to more than 30,000 children and families. Previously, he worked for The Walt Disney Company, where he ran the global mobile games business after starting it in Europe.

Laura S. Hamilton is a senior associate at the National Center for the Improvement of Educational Assessment, where she collaborates with states, districts, and nonprofit organizations on the design and implementation of assessment policies and practices. She is especially interested in supporting the development and implementation of large-scale and classroom assessment systems that measure students' civic readiness, and she is co-editing a volume on assessing civic learning and engagement. Her previous roles include senior director at American Institutes for Research, associate vice president in the Research and Measurement Sciences area at ETS, distinguished chair in learning and assessment at RAND, and co-director of RAND's nationally representative educator survey panels. Hamilton regularly serves on expert committees and panels including the Joint Committee to revise the AERA/APA/NCME Standards for Educational and Psychological Testing, multiple National Academies of Sciences, Engineering, and Medicine committees, and technical advisory committees for state assessment programs. She's also held editorial roles with several journals. She is a fellow of the American Educational Research Association and received the Joseph A. Zins Distinguished Scholar Award for Social and Emotional Learning Research. Hamilton earned a Ph.D. in educational psychology and an M.S. in statistics from Stanford University.

Emily C. Hanno is a Senior Research Associate at MDRC where she is Project Director and co-Principal Investigator of the Measures for Early Success Initiative. Hanno's research, which is grounded in her experiences as a Head Start teacher and instructional coach, focuses on understanding how early education and care innovations, programs, and policies can support children, families, and communities.

John Hattie is Emeritus Laureate Professor at the Melbourne Graduate School of Education at the University of Melbourne, Chief Academic Advisor for Corwin, i-Ready Technical Advisor, and co-director of the Hattie Family Foundation. His career was as a measurement and statistics researcher and teacher, and his more recent research, better known as Visible Learning, is a culmination of nearly 30 years synthesizing more than 2,500 meta-analyses comprising more than 140,000 studies involving over 300 million students around the world.

Dr. Norris M. Haynes is a Professor in the Educational Leadership Department at Southern Connecticut State University. He founded and directed the Center for Community and School Action Research (CCSAR) and served as Chairperson of the Counseling and School Psychology Department. Dr. Haynes is a Clinical faculty member at the Yale University School of Medicine Child Study Center and where he has been an Associate Professor and Director of Research for the Yale University Comer School Development Program. He earned his Ph.D. in Educational Psychology and an M.B.A. with a focus on health services administration from Howard University. Haynes is a licensed Psychologist, Fellow of the American Psychological Association, and Diplomate in the International Academy for Behavioral Medicine, Counseling, and Psychotherapy. His research interests include social-emotional learning, school climate, resilience, and academic achievement. Dr. Haynes has authored numerous articles, books, and evaluation reports. He is a founding leadership team member of the Collaborative for Academic and Social Learning (CASEL) and researcher with Social Emotional and Character Development (SECD). He has worked with educational and psychological entities to enhance school practices. Dr. Haynes has been involved in national research initiatives, including studies on youth violence, social and emotional learning, and the Harlem Children's Zone (HCZ) programs.

JoAnn Hsueh is currently Vice President of Program and Communications at the Foundation for Child Development and co-Principal Investigator and Senior Advisor for the Measures for Early Success Initiative. Trained as a developmental scientist, Hsueh has broad interests in studying the impact and implementation of social, economic, and educational policies and programs that influence family and child well-being.

Kristen Huff, M.Ed**., Ed.D.,** currently serves as the Head of Measurement at Curriculum Associates, where she leads a team of assessment designers, psychometricians, and researchers in the development of online assessments integrated with personalized learning and teacher-led instruction. Prior to this role, she served as the Senior Fellow for the New York State Education Department as well as serving in leadership roles with several major assessment companies. Dr. Huff has deep expertise in K–12 large scale assessment, and has presented and published consistently in educational measurement conferences and publications for over 25 years. She served previously as a technical advisor for the 2026 NAEP Frameworks in Reading and Mathematics and as the inaugural Co-Chair of the NCME Task Force on Classroom Assessment 2016-2020. She was named as recipient of the 2021 Career Achievement Award from the Association of Test Publishers, and now serves as the NCME Representative to the Management Committee for the revision of the 2014 Joint Standards for Educational and Psychological Testing, published by AERA, APA, and NCME. Dr. Huff is first author of the forthcoming *Educational Measurement, 5th Edition* (Oxford University Press), and Designing and Developing Educational Assessments (Huff, Nichols, and Schneider).

Diana Hughes is Head of Product at Relay Graduate School of Education. She is an experienced practitioner of game design and personalized learning. As VP of Learning Science and Design at Age of Learning, Inc., Diana led the development of Age of Learning's science-backed, evidence-centered programs, My Math Academy, My Reading Academy, and My Reading Academy Español. With three patents in personalized learning technologies to her name, Diana is known for her innovative and effective contributions to digital education methodologies. Her work, underpinned by a profound commitment to student-centric design and efficacy, exemplifies her dedication to providing equitable, effective, and engaging learning experiences for children globally. Diana's past work includes an empathy game for children on the autism spectrum, a graphics-free game for blind and low-vision players, and soft skills training games for the United States Military. She holds an MFA in Game and Interactive Design from the University of Southern California and a BS in Multimedia from Bradley University.

Gerunda B. Hughes is Professor Emerita, Howard University. During her tenure at the University, Dr. Hughes served as Director of the Office of Institutional Assessment & Evaluation and Professor of Mathematics Education. As Director, she oversaw the collection and analyses of student learning and other institutional-level data. She also served as coordinator of secondary education programs and taught courses in mathematics, mathematics pedagogy, assessment and measurement, and research methodology. Dr. Hughes served as Principal Investigator of the "Classroom Assessment Project" at Howard University's Center for Research on the Education of Students Placed at Risk (CRESPAR). She was an inaugural member of the Board of Directors of the Howard University Middle School for Mathematics and Science. Dr. Hughes has served as Co-Editor-in-Chief of the Journal of Negro Education; Associate Editor of Review of Educational Research; and a member of the editorial boards of the American Educational Research Journal and the Mathematics Teaching-Research Journal. She currently serves on technical advisory committees for national, state, and professional testing and assessment organizations. Dr. Hughes earned a B.S. in mathematics from the University of Rhode Island, a M.A. in mathematics from the University of Maryland-College Park, and a Ph.D. in educational psychology from Howard University.

Neal Kingston, Ph.D., is University Distinguished Professor in the Department of Educational Psychology at the University of Kansas, Director of the Achievement and Assessment Institute (AAI), and Vice Provost for Jayhawk Global and Competency-Based Education. His research focuses on large-scale assessment, with particular emphasis on how it can better support student learning through the use of learning maps and diagnostic classification models. Current interests include games-based assessment, personalizing assessments to improve student engagement, and the creation of more agile test development approaches. Dr. Kingston has served as principal investigator or co-principal investigator for over 250 research grants. Of particular note was the Dynamic Learning Maps Alternate Assessment grant from the US Department of Education, which was at that time was the largest grant in KU history and which currently serves 23 state departments of education. Other important testing projects include the Kansas Assessment Program, Project Lead The Way, and Adaptive Reading Motivation Measures. He is known internationally for his work on large-scale assessment, formative assessment, and learning maps. He has served as a consultant or advisor for organizations such as the AT&T, College Board, Department of Defense Advisory Committee on Military Personnel Testing, Edvantia, General Equivalency Diploma (GED), Kaplan, King Fahd University of Petroleum and Minerals, Merrill Lynch, National Council on Disability, Qeyas (Saudi Arabian National Center for Assessment in Higher Education), the state of New Hampshire, the state of Utah, the U.S. Department of Education, and Western Governors University.

Geoffrey T. LaFlair is a Principal Assessment Scientist at Duolingo where he co-leads Assessment Research and Development for the Duolingo English Test. He holds an MA in TESOL from Central Michigan University and a Ph.D. in Applied Linguistics from Northern Arizona University. Prior to joining Duolingo, he was an Assistant Professor in the Department of Second Language Studies at the University of Hawai'i at Mānoa and the Director of Assessment in the Center for ESL at the University of Kentucky. His research interests are situated at the intersection of language assessment, psychometrics, and natural language processing, focusing on the application of research from these fields in researching and developing operational language assessments.

Carol D. Lee is the Edwina S. Tarry Professor Emeritus of Education in the School of Education and Social Policy and in African-American Studies at Northwestern University, and the President of the National Academy of Education. She is currently Chairman of the National Board of Education Sciences. She is a past president of the American Educational Research Association (AERA) and past president of the National Conference on Research in Language and Literacy. She is a member of the American Academy of Arts and Sciences and a fellow of the American Educational Research Association. She has won numerous awards and honors, including the McGraw Prize in Education. Her research addresses cultural supports for learning that include a broad ecological focus, integrating learning sciences and human development framing, with attention to language and literacy and African American youth. She is the author or co-editor of eleven books, monographs and special issues, including co-editing The Handbook of Cultural Foundations of Learning, and has published over 108 journal articles and book or handbook chapters in the field of education. She has also worked as an English Language Arts teacher and a primary grade teacher. She is a founder of four African-centered schools.

Paul G. LeMahieu is Senior Fellow at the Carnegie Foundation for the Advancement of Teaching and graduate faculty in education, University of Hawai'i at Mānoa. LeMahieu served as Superintendent of Education for the State of Hawai'i, serving 190,000 students. Prior to that, he was Undersecretary for Education Policy and Research for the State of Delaware. He has been President of the National Association of Test Directors and Vice President of the American Educational Research Association. He served on the National Academy of Sciences' Board on International Comparative Studies in Education, Mathematical Sciences Board, National Board on Testing Policy, and the National Board on Professional Teaching Standards. His professional interests focus on the adaptation of improvement science methodologies for application in networks in education. He is a co-author of the book *Learning to Improve: How America's Schools Can Get Better at Getting Better* (2015), and lead editor of the volume *Working to Improve: Seven Approaches to Improvement Science in Education* (2017). His most recent book is entitled *Measuring to Improve: Practical Measurement to Support Continuous Improvement in Education* (2025). Paul has a Ph.D. from the University of Pittsburgh, an M.Ed. from Harvard University, and an A.B. from Yale College.

Richard M. Lerner is the Bergstrom Chair in Applied Developmental Science and the Director of the Institute for Applied Research in Youth Development at Tufts University. He went from kindergarten through Ph.D. within the New York City public schools, completing his doctorate at the City University of New York in 1971 in developmental psychology. Lerner has more than 800 scholarly publications, including 90 authored or edited books. He was the founding editor of the Journal of Research on Adolescence and of Applied Developmental Science. He is currently the Editor of Review of General Psychology, the flagship journal of Division 1 of the American Psychological Association (APA). Lerner was a 1980-81 fellow at the Center for Advanced Study in the Behavioral Sciences and is a fellow of the American Association for the Advancement of Science, the APA, and the Association for Psychological Science (APS). He is the recipient of several awards for his career achievements: The SRA John P. Hill Memorial Award for Life-Time Outstanding work (2010); the APA Division 7 Urie Bronfenbrenner Award for Lifetime Contribution to Developmental Psychology in the Service of Science and Society (2013); the APA Gold Medal for Life Achievement in the Application of Psychology (2014); the APA Division 1 Ernest R. Hilgard Lifetime Achievement Award for distinguished career contributions to general psychology (2015); the ISSBD Award for the Applications of Behavioral Development Theory and Research (2016); the SRCD Distinguished Contributions to Public Policy and Practice in Child Development Award (2017); the APS James McKeen Cattell Fellow Award winner for lifetime outstanding contributions to applied psychological research (2020); and the SSHD Distinguished Lifetime Career Award (2021). Lerner served on the Board of Directors of the Military Child Education Coalition for 10 years and still serves on their Scientific Advisory Board. In February 2023, Pope Francis reappointed Lerner to a second five-year term as a Corresponding Member of the Pontifical Academy for Life.

Lei Liu is a Research Director leading the K–12 research team at ETS. She is also an Adjunct Professor at the University of Pennsylvania. Her research interests lie at the intersection of science learning and assessment, learning sciences, and educational technology. She has led multiple federal grants to develop transformative innovations for STEM learning, including topics on learning progressions, AI-supported assessment tools, and virtual labs. She has produced over 70 peer-reviewed publications. She is a member of the editorial board of Instructional Science and has served as a reviewer for multiple international conferences, journals, and NSF merit reviews. In addition to her lead role in research, Dr. Liu has also been a key contributor to support various operational works at ETS including the California State Assessment programs, and NAEP science and mathematics programs. She earned a Ph.D. in educational psychology with a focus on learning sciences and educational technology from Rutgers University.

Ou Lydia Liu, Associate Vice President of Research at ETS, is a globally recognized expert in assessment of critical skills and competencies in higher education and workforce. She has also managed large-scale grants awarded by government and private funding agencies in the U.S. and international countries including India, China, and Korea. Dr. Liu has authored and coauthored over 100 peer-reviewed journal articles, research reports, and book chapters in the fields of applied measurement, higher education, and science assessment. Her research appeared in Science, Nature Human Behavior, Educational Researcher, and other influential outlets. She delivered over 100 invited seminars and peer-reviewed conference presentations domestically and internationally. Dr. Liu was inducted as an AERA Fellow in 2023, and received the 2019 Robert Linn Memorial Lecture Award, and the 2011 National Council on Measurement in Education Jason Millman Promising Measurement Scholar Award in recognition of her original and extensive research in learning outcomes assessment in higher education and K–12 science assessment. Dr. Liu holds a doctorate in Quantitative Methods and Evaluation from the University of California, Berkeley.

Silvia Lovato is head of Learning & Research at PBS KIDS, where she leads the team responsible for PBS KIDS curriculum development, research and evaluation, and early childhood education strategy. Previously, she worked at PBS KIDS from 2000 to 2014 as a Content Manager and Senior Product Director, managing the production of interactive features for PBS KIDS digital platforms, especially games. A seasoned children's media professional and researcher who is passionate about how media can help kids learn, Silvia holds a Ph.D. in Media, Technology and Society from Northwestern University. Her dissertation, titled "Hey Google, Do Unicorns Exist?", explored how children use AI-based conversational agents such as the Google Assistant to seek answers to their many questions. She holds certificates in Cognitive Science and Management for Scientists and Engineers.

Dr. Temple S. Lovelace is the Executive Director of Assessment for Good (AFG), an inclusive R&D program supported by the Advanced Education Research and Development Fund (AERDF). AFG focuses on creating new assessment tools that explore how we recognize and maximize each student's potential as they leverage a unique set of skills to power their personal learning journey. In 2018, Temple launched a groundbreaking cooperative incubator in the School of Education at Duquesne University. There, she developed an innovative research and development methodology now being implemented by organizations across the United States. Her successful community-engaged programs—Youth Leading Change, Education Uncontained, and Girlhood Rising—have empowered educators and students to conduct localized R&D that bridges innovation and effective learning practices. Now, as a visiting scholar at the Gordon Institute for Advanced Study at Teachers College, Columbia University, Temple's research explores the role of context-capable assessment and learning so that we can understand the fullness of how learners explore their world and translate that to more modernized understandings of child development. A respected voice in educational innovation, Temple has published extensively on assessment design and student-centered learning approaches with the hope that educators, caregivers, and even learners themselves can co-create a future where all learners thrive.

Susan Lyons, Ph.D., works to transform traditional assessment systems to better serve the needs of students, educators, and the public. As the Principal Consultant at Lyons Assessment Consulting, Susan partners with innovators to advance theory and practice in educational measurement. Susan holds a bachelor's degree in Mathematics and Math Education from Boston University and served as a math educator before pursuing her graduate work. She received her master's and Ph.D. in Educational Psychology with a focus on Research, Evaluation, Measurement and Statistics from the University of Kansas. Susan is the co-founder of Women in Measurement, a nonprofit organization dedicated to advancing gender and racial equity in the field. Since its launch, she has served as the organization's Executive Director, ushering it through the start-up phase to its now prominent position as a fixture within the measurement community, offering support for more than a thousand women in our field.

Scott F. Marion, Ph.D., is a principal learning associate at the National Center for the Improvement of Educational Assessment. He is a national leader in conceptualizing and designing innovative and balanced assessment systems to support instructional and other critical uses. He has also led extensive work across the country to design and implement school accountability systems. Scott is an elected member of the National Academy of Education and is one of three measurement specialists on the National Assessment Governing Board, which oversees the National Assessment of Educational Progress. He coordinates and/or serves on 10 state or district technical advisory committees for assessment and accountability. He has served on multiple National Research Council committees, including those that provided guidance for next-generation science assessments, investigated the issues and challenges of incorporating value-added measures in educational accountability systems, and outlined best practices in state assessment systems. Scott is a co-author of the validity chapter in the 5th edition of Educational Measurement, a co-editor of the National Academy of Education's Reimagining Balanced Assessment, and a co-author of Instructionally Useful Assessment. He has published dozens of articles in peer-reviewed journals and edited volumes, and he regularly presents his work at the national conferences of the American Educational Research Association, National Council on Measurement in Education, and the Council of Chief State School Officers. Scott earned a Ph.D. from the University of Colorado Boulder with a concentration in measurement and evaluation.

Kimberly McIntee centers social (in)justice in developing equitable academic and assessment strategies and improving how results are created and shared. Her research examines testing procedures, assessment theories, and critiques of the harm curricula and assessments can cause individuals and society, with the goal of transforming traditional testing into meaningful practices that support teaching and learning. Growing up in a multiracial, multilingual environment pushed McIntee to constantly reflect on her identity and experiences across psychological, physical, and social dimensions. McIntee's earliest school memories involve navigating between worlds. This divide deepened when she and a few other minoritized peers were placed in classes where, despite attending predominantly Black schools, the majority of students became invisible in halls saturated with unfamiliar white faces. Such segregation often stemmed from curricula and assessments designed without accounting for diverse learners, particularly those least prepared by inequitable systems. Recognizing these hidden patterns of separation, McIntee advocates for schools where students' identities do not isolate them and where statistics do not dictate resources. She believes that through intentional research and just assessment design, academic and social spaces—long marked by inequity—can be reshaped into sites of empowerment.

Maxine McKinney de Royston is the Dean of Faculty at the Erikson Institute. Dr. McKinney de Royston's research and teaching examine how educators' political clarity can be reflected in their pedagogical practices in ways that support the intellectual thriving and holistic well-being of racially and economically minoritized learners. She is a co-editor, along with Na'ilah Suad Nasir, Erikson's Trustee Carol Lee, and Roy Pea, of the Handbook of the Cultural Foundations of Learning; free access: https://doi.org/10.4324/9780203774977. In addition to numerous peer-reviewed articles, chapters, and other publications and presentations, Dr. McKinney de Royston has served as Associate Editor of the American Educational Research Journal, Co-Chair of the Wallace Foundation Emerging Scholars Committee, and Advisor to the Wisconsin Department of Public Instruction, Family, Youth, & Community Advisory Council. She is a member of several professional learned societies, including the American Educational Research Association (AERA), the International Society of the Learning Sciences, the National Association for Multicultural Education, and the National Council of Black Studies.

Elizabeth Mokyr Horner is a Senior Program Officer at the Gates Foundation, which provided grant funding to support MDRC's Measures for Early Success Initiative. Dr. Mokyr Horner worked in partnership with MDRC to develop the approach to codesign described in this chapter. She has spent the last 15+ years across academic, non-profit, government, and foundation sectors supporting and evaluating evidence-based interventions designed to enhance educational outcomes, economic opportunity, and improved overall quality of life.

Orrin T. Murray, Ph.D., a learning scientist, is principal of the Wallis Research Group. Through Wallis Research Group, he has advised leading institutions, providing research, equity-driven program evaluations, and AI-based insights to shape social impact initiatives. He has been a workshop leader and mentor/coach, building evaluation skills and capacity in community-based organizations in Chicago and Cincinnati. As a Principal Researcher at the American Institutes for Research, he led national studies on education equity, civic education, AI-driven learning, and workforce development, ensuring that data-driven insights lead to real-world improvements. His thought leadership has shaped policy decisions, education strategies, and AI integration in learning, making him a trusted advisor to policymakers, school districts, and nonprofit organizations. At the University of Chicago's Urban Education Institute, he led a digital foundry responsible for designing and launching research-based tools to improve high school and college completion rates. Orrin's expertise extends into culturally responsive teaching, having contributed to "Culture in Our Classrooms," a documentary viewing guide on fostering belonging and inclusion in education. He is also a recognized voice in AI and education research, co-authoring "Principles to Guide Artificial Intelligence in Education Research", which outlines ethical considerations and bias mitigation in AI applications.

Na'ilah Suad Nasir is the sixth President of the Spencer Foundation, which funds education research nationally. Prior to joining Spencer, she held a faculty appointment in Education and African American Studies at the University of California, Berkeley where she also served as the chair of African American Studies, then later as the Vice Chancellor for Equity and Inclusion. Her scholarship focuses on race, culture, and learning, and how what we know about learning has implications for how we design schools for equity. In her foundation work, she has worked to bring a deep equity lens to grantmaking, and has spearheaded innovative funding opportunities rooted in the promise of research to support more equitable education systems. She is a member of the American Academy of Arts & Sciences and the National Academy of Education, and is a Fellow of the American Educational Research Association. She is a Past President of the American Educational Research Association and serves on the board of Sage Publications, the National Equity Project, and the UC Berkeley Board of Visitors.

Michelle Odemwingie is the chief executive officer at Achievement Network. Michelle joined ANet nearly a decade ago as a coach and has since held roles as chief of school and system services and chief of staff, among others. This includes spearheading ANet's Breakthrough Results Fund in partnership with five school districts across the country. Through her work at ANet and in her local community, Michelle maintains a deep personal commitment to educational equity and ensuring all students are able to learn and thrive. A recognized strategic advisor and policy advocate for the future of assessments, she plays a key role in shaping the national conversation around instructional improvement. Michelle actively engages in education policy and system-level transformation, advising districts, policymakers, and nonprofit leaders on instructional strategy, assessment innovation, and equitable access to high-quality materials. Prior to joining ANet, she spearheaded the ThinkMath team in California and DC, supporting instructional leaders around math enrichment and intervention programs, as well as supporting secondary math teachers through TNTP and Teach for America. Michelle began her career as an educator teaching math in the District of Columbia and is a graduate of Stanford University.

Maria Elena Oliveri is a Research Associate Professor of Engineering Education at Purdue University, working on the SCALE program. She is dedicated to developing innovative and equitable assessment approaches that prepare learners for professional practice. Her research focuses on improving assessment methods in engineering learning contexts, with particular attention to fairness, culturally and linguistically relevant assessment, assessing complex engineering competencies, and aligning assessments with evolving workforce needs. She has extensive expertise in the development of simulations, performance-based assessments, and the assessment of complex professional skills. She has played a leading role in shaping international assessment standards and best practices. She served as Chair for the International Test Commission's (ITC) Guidelines for the Fair and Valid Assessment of Linguistically Diverse Populations and as a steering committee member for the ITC Technology-Based Assessment Guidelines. She has authored various guidelines and standards in the field of assessment and has published over 100 peer-reviewed journal articles and conference papers. She is a multilingual researcher and speaks Spanish, French, and Italian. Her research continues to advance equity and effectiveness in education and workplace readiness.

Saskia Op den Bosch is co-founder of RevX, where she leads R&D strategy and spearheads the development of our innovative assessment system. She brings 14 years of experience as an educational researcher, strategist, and peer-reviewed author, creating environments that foster a strong sense of self and community, intellectual growth, and real-world impact. Previously, she led R&D for Getting Ready for School, integrating SEL into early literacy across NYC Head Start centers, and coached grantees at Character Lab on translating research into classroom practice. As Partner of R&D at Transcend, she built the R&D blueprint that secured large-scale federal funding for the Whole Child Model. Saskia holds a B.S. in Psychology from Carnegie Mellon and an M.A. in Quantitative Methods from Columbia. Committed to reimagining assessment as a catalyst for growth, she ensures learning environments evolve alongside young people—equipping learners to step into their purpose and create meaningful impact.

Dr. V. Elizabeth Owen is an expert in game-based learning analytics, with over 20 years experience in the learning sciences and education. At Age of Learning, she specializes in optimizing adaptive learning systems through applied AI and machine learning. Previously, she worked as a researcher and data scientist with Google, GlassLab Games at Electronic Arts, Inc. (EA) and LRNG by Collective Shift, after earning a Ph.D. in Digital Media (Learning Analytics focus) from the University of Wisconsin-Madison. Dr. Owen's doctoral work was based at the Games+Learning+Society (GLS) center, which launched collaborations with EA, Zynga, and PopCap Games using game-based Educational Data Mining. Dr. Owen spent a decade as a K–12 educator and was a founding teacher at the Los Angeles Academy of Arts and Enterprise charter school. She holds a BA from Claremont McKenna College.

Trevor Packer is the head of College Board's Advanced Placement Program. In rigorous classes that range from calculus to studio art, Advanced Placement provides high-quality coursework and the opportunity for college credit to more than 3 million students every year. With a deep love for literature, Trevor spent his time prior to the College Board working in academia. He has taught composition and literature at the City University of New York and Brigham Young University.

Roy Pea is David Jacks Professor of Education & Learning Sciences at Stanford University, Graduate School of Education, and Computer Science (Courtesy). His extensive publications in the learning sciences focus on advancing theories, research, tools and social practices of technology-enhanced learning of complex domains. He founded and directs Stanford's Ph.D. program in Learning Sciences and Technology Design. He is a Fellow of the American Academy of Arts and Sciences, National Academy of Education, Association for Psychological Science, the American Educational Research Association, and The International Society for the Learning Sciences. His most recent books include Learning Analytics in Education (2018), The Routledge Handbook of the Cultural Foundations of Learning (2020), and AI in Education: Designing the Future (2023). He is co-author of the National Academy of Sciences books: How People Learn (2000), and Planning for Two Transformations in Education and Learning Technology (2003). His most recent research involves studies of appropriate roles for Generative AI in augmenting writing and its development, computer science education, virtual reality storytelling, and culturally responsive science learning with augmented reality. In 2018 he received an Honorary Doctorate from The Open University. He won the McGraw Prize for Learning Sciences Research in 2022.

James W. Pellegrino is Emeritus Professor of Psychology and Learning Sciences and Founding co-director of the Learning Sciences Research Institute at the University of Illinois Chicago. His research and development interests focus on children and adults thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 350 books, chapters, and articles on cognition, instruction, and assessment. His education research has been funded by the National Science Foundation, the Institute of Education Sciences, and private foundations. As Chair or Co-Chair of several National Academy of Sciences study committees he co-edited major synthesis reports on teaching, learning, and assessment, including *Knowing What Students Know: The Science and Design of Educational Assessment*. He previously served on the Board on Testing and Assessment of the National Research Council and is a lifetime member of both the National Academy of Education and the American Academy of Arts and Sciences. His service includes the Technical Advisory Committees of several states and consortia, as well as those of the College Board, ETS, OECD, and the National Center on Education and the Economy. He currently serves on the NAEP Validity Studies Panel and ETS' Visiting Panel on Research.

Mario Piacentini is a Senior Analyst in the Programme for International Student Assessment (PISA) at the Organisation for Economic Co-operation and Development (OECD). An expert in measurement, Mario leads the work on the PISA innovative assessments and the broader PISA Research & Development Programme. He works with international experts to design assessments of 21st century competences. His projects aim to expand the metrics we use to define successful education systems. He is one of the authors of the Global Competence (PISA 2018) and Creative Thinking (PISA 2022) assessment frameworks, and he is currently leading the development of the PISA 2025 assessment of Learning in the Digital World and PISA 2029 assessment of Media and AI Literacy. He also coordinates the development of an open-source platform to support the use of technology-enhanced, formative assessments in the classroom. Before joining PISA, he worked for the Public Governance and the Statistics Directorates of the OECD, the University of Geneva, the World Bank and the Swiss Cooperation. He has authored several peer-reviewed articles and reports and was co-editor of the OECD publication on Innovating Assessments to Measure and Support Complex Skills. Mario holds a Ph.D. in economics from the University of Geneva.

Mya Poe is Professor of English at Northeastern University. Her research focuses on writing assessment and writing development with particular attention to justice and fairness. For more than 20 years she has advocated against assessment practices that are based on weak construct models and that result in unnecessary barriers for students. She has published five books, including Learning to Communicate in Science and Engineering: Case Studies from MIT (CCCC 2012 Advancement of Knowledge Award); Race and Writing Assessment (CCCC 2014 Outstanding Book of the Year); Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education(CWPA 2022 Book of the Year); and Rethinking Multilingual Writers in Higher Education: An Institutional Case Study. In addition to teaching undergraduate courses on writing research methods and scientific writing, she also teaches graduate courses on writing assessment and the teaching of writing. Her teaching and service have been recognized with the Northeastern University Teaching Excellence Award and the MIT Infinite Mile Award for Continued Outstanding Service and Innovative Teaching. She has directed writing programs at MIT and Northeastern University and has worked extensively with faculty across the U.S. to improve the teaching of writing. She is co-editor of the international writing research journal Written Communication.

Ximena A. Portilla is a Senior Research Associate at MDRC where she serves as Content Lead for the Measures for Early Success Initiative, shaping a vision for the assessment content covered by tools coming out of the initiative and connecting assessment developers to supports to ensure content is aligned with developmental science. Portilla is a developmental scientist whose research over the last 20 years has focused on a range of topics in the preschool and kindergarten years, including home visiting, school readiness, and classroom supports for early educators.

Dr. Elizabeth J. K. H. Redman is a Research Scientist specializing in technology and assessment at the National Center for Research in Evaluation, Standards, and Student Testing (CRESST). Her primary research interests include STEM education, educational games, and assessment design. Her recent research focus has been on incorporating assessment capabilities into educational games, including SEL and STEM games. She has experience running observational classroom studies, RCTs and evaluations of educational games.

Jeremy D. Roberts is Senior Director of Learning Technology for PBS KIDS, where he works closely with award-winning series such as Curious George, Molly of Denali, Work it Out Wombats!, and Lyla in the Loop to deliver innovative, educational, multi-platform media experiences to kids aged 2-8. Roberts' work focuses on demonstrating and optimizing the impact produced by PBS KIDS media at scale. One of Roberts' core initiatives is the PBS KIDS Learning Analytics research program, which uses safe anonymous gameplay data, analytics, statistical modeling, research, and AB testing, to systematically discover game design principles that best balance reach, engagement, and learning effectiveness. Roberts' work helps PBS KIDS improve its overall impact by feeding relevant insights directly into the design, production, packaging, and distribution of PBS KIDS media. Over the decades, Roberts has cultivated a deep strategic understanding of technology, and the fast-evolving nature of the media, entertainment, and learning landscapes. A physicist by training, Roberts' passion for discovery and innovation has driven his extensive involvement with leading-edge technologies, and continues to define his work as an executive, leader, strategist, and systems engineer. To keep things interesting, Roberts plays trombone with D.C. soul, ska, and reggae band The Pietasters.

Dr. Mary-Celeste Schreuder is the Director of Literacy at the Achievement Network (ANet), where she leads ANet's national rollout of the Rapid Online Assessment of Reading (ROAR) in collaboration with Stanford University. With 20+ years in education, including roles as a secondary ELA teacher, professor of teacher education, and literacy strategist, Mary has built deep expertise in adolescent literacy, assessment strategy, and writing pedagogy. She designs tools, leads professional learning, and equips coaches and system leaders to support striving readers through research-based, equity-centered solutions. Her scholarship has been published in journals like the *Journal of Adolescent & Adult Literacy*, and she holds a Ph.D. in Literacy, Language, and Culture from Clemson University.

David Sherer is Director, Future of Assessment, at the Carnegie Foundation. In this role, he leads the Skills for the Future initiative, in collaboration with colleagues at ETS, to create a robust, scalable suite of assessment and analytic tools that captures the full range of skills required for students to succeed in K–12, post-secondary education and beyond. David coaches educational leaders in the use of evidence in the improvement process, the development of indicators and measures, and the assessment of organizational health. He holds a master's degree and a doctorate (Ed.D.) from the Harvard Graduate School of Education.

Stephen G. Sireci, Ph.D., is Distinguished Professor and Executive Director of the Center for Educational Assessment in the College of Education, University of Massachusetts Amherst. He earned his Ph.D. in psychometrics from Fordham University and his master and bachelor degrees in psychology from Loyola College Maryland. Before UMass, he was Senior Psychometrician at GED Testing Service, Psychometrician for the CPA Exam and Research Supervisor of Testing for the Newark NJ Board of Education. He is known for his research in validity and fairness of educational tests, and for innovations in test development. He currently serves/has served on several advisory boards including the National Board of Professional Teaching Standards, Duolingo English Test, and technical advisory committees for Florida, Maryland, New Hampshire, New York, Montana, Puerto Rico, and Texas. He is a Fellow of American Educational Research Association and of Division 5 of American Psychological Association, and a lifetime member of the National Academy of Education. He is a past President of International Test Commission, Northeastern Educational Research Association, and National Council on Measurement in Education. His UMass honors include School of Education's Outstanding Teacher Award, Conti Faculty Fellowship, Public Engagement Fellowship, Outstanding Accomplishments in Research and Creative Activity Award, and the Chancellor's Medal. He also received the Messick Memorial Lecture Award from Educational Testing Service/International Language Testing Association. He serves on several editorial boards including Applied Measurement in Education, Educational Assessment, Educational Measurement: Issues and Practice, Educational and Psychological Measurement, Practical Assessment Research and Evaluation, and Psicothema.

Dr. Erica Snow is the Senior Director of People Science and Analytics and Early Career Recruiting at Roblox. Previously, she was Director of Learning and Data Science at Imbellus, a game-based assessment startup acquired by Roblox. She also worked at SRI international as the Lead Learning Analytics Scientist before joining Imbellus. Dr. Snow has over a decade of experience evaluating the implementation and impact of a variety of educational technologies(i.e., ITSs, MOOCs, LMS, and blended learning courses) within K–12, postsecondary education, and workforce training. Her work has been presented both domestically and internationally to both scientific and non-scientific colleagues and has been published in over 70 peer-reviewed publications. She holds a Ph.D. and MA in Cognitive Science from Arizona State University and a BA in Psychology from Ball State University.

Rebecca A. Stone-Danahy has served as College Board's Director of AP Art and Design since 2020, where she has spearheaded initiatives to support course growth and advocacy ensuring access to inquiry-based art education through assessment practices. She also led the transformation of a physical to digital annual AP Art and Design exhibit, enhancing the visibility of diverse and high-quality student artworks. Rebecca's leadership in K–12 education spans roles from visual arts educator to fine arts administrator where she focused on inquiry-based visual art pedagogy, curriculum design, fine arts programming, and teacher mentorship. She is a strong proponent of integrating technology into education and was pivotal in launching one of the first online distance learning programs and museum collaborations between the North Carolina Virtual Public Schools and the North Carolina Museum of Art. Rebecca holds an MA in Art Education from Miami University in Oxford, Ohio, an M.Ed. in Secondary School Administration and an Ed.S. in Educational Leadership-School Superintendent from The Citadel in Charleston, SC, and an Ed.D. in Educational Systems Improvement Science from Clemson University in Clemson, SC. Rebecca's dissertation focus aimed to improve access and equity to inquiry-based visual art education for Title I school students in South Carolina.

Rebecca Sutherland, Ed.D., is the Associate Director of Research at Reading Reimagined, a funded program of the Advanced Education Research and Development Fund, where she leads a portfolio of research projects investigating the root causes of reading struggles among older students and instructional resources designed to address them. Rebecca has worked with K–12 public education data for over two decades to generate actionable knowledge for state and local agencies, and nonprofit organizations. She has taught ESL and reading in public schools in Japan and New York, and adult literacy in New York and Massachusetts. Rebecca holds a doctorate in Human Development and Psychology from the Harvard Graduate School of Education, a masters degree in Educational Psychology from the New York University Steinhardt School of Education, and a B.A. in history from Barnard College.

Natalya Tabony is Executive Director of AP Strategy and Analytics at the College Board. She leads a team focused on shaping program and product strategies that help more students access—and succeed in—Advanced Placement. Her work centers on using data and research to guide thoughtful decisions about how to strengthen the AP program and ensure it meets the needs of students and schools. Natalya began her career as a consultant with Parthenon-EY's education practice, where she worked on strategy and growth projects for school systems, universities, and philanthropic foundations. She later served as Director of Operations at a middle school in the Uncommon Schools network in Brooklyn, overseeing all aspects of daily operations. Across roles, she's been drawn to questions about how to improve schools and create more moments where students can discover what they're capable of. She emigrated from Russia to the U.S. as a child and grew up believing in the power of education to shape opportunity. Natalya holds a BA from Dartmouth College and an M.B.A. from the Kellogg School of Management. She lives in New York City with her husband and two young children.

Carrie Townley-Flores is the Director of Research and Partnerships for the Rapid Online Assessment of Reading (ROAR) at Stanford University. She holds a Ph.D. in Education Policy from Stanford. Her research focuses on reading assessment and related policies and practices that mitigate racial, ethnic, and economic inequality in the U.S. She joined the ROAR project with extensive experience working with schools, both in the classroom and in academic research-practice partnerships. Carrie taught English Language Arts at secondary schools in Michigan and New Hampshire and a primary school in Helsinki, Finland. She holds a B.A. in English and Education from University of Michigan.

Eric M. Tucker is the President and CEO of the Study Group, which exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy. He has served as President of Equity by Design, Superintendent and Executive Director of Brooklyn Laboratory Charter Schools, CEO of Friends of Brooklyn LAB, Cofounder of Educating All Learners Alliance, Executive Director of InnovateEDU, director at the Federal Reserve Bank of New York, and Cofounder and Chief Academic Officer of the National Association for Urban Debate Leagues. As an entrepreneurial, strategic, and impact-focused leader, Eric has over 25 years of experience building catalytic partnerships in education, securing over $300 million of investments for enterprises and initiatives that have transformed outcomes for learners and educators. Eric has expertise in measurement and assessment system innovation, participatory and advanced R&D, analytics, and human infrastructures for improvement and co-edited *The Sage Handbook of Measurement*. He earned a doctorate and a masters of science in measurement sciences from the University of Oxford and bachelors degrees from Brown University. Eric served as an ETS MacArthur Foundation Fellow with the Gordon Commission on the Future of Assessment in Education. He served as a Senior Research Scientist at the University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

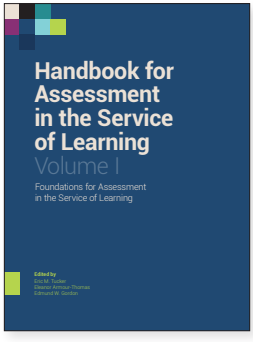
Alina A. von Davier is a researcher, innovator, and an executive leader with over 20 years of experience in EdTech and in the assessment industries. She is the Chief of Assessment at Duolingo, leading the Duolingo English Test research and development area. She is the Founder and CEO of EdAstra Tech. She is an American Educational Research Association (AERA) Fellow and serves as an Honorary Research Fellow at University of Oxford, and a Senior Research Fellow Carnegie Mellon University. Her research spans computational psychometrics, machine learning, and education. Dr. von Davier's work has been widely recognized in the academic community. She received the Brad Hanson award twice from National Council on Measurement in Education (NCME) for her pioneering work on computational psychometrics, and her work on adaptive testing. She received ATP's Career Award for her contributions to assessment. She was a finalist for the Innovator award from the EdTech Digest. The AERA awarded her the Division D Signification Contribution Educational Measurement and Research Methodology Award for her publications "Computerized Multistage Testing: Theory and Applications" (2014) and an edited volume on test equating, "Statistical Models for Test Equating, Scaling, and Linking" (2011).

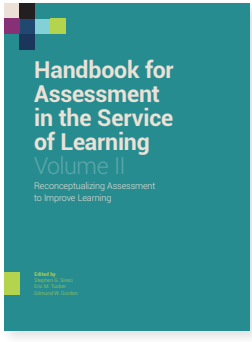
Kevin Yancey is a Senior Staff AI Researcher at Duolingo, leading the engineering and AI functions for Research & Development on the Duolingo English Test.As an expert software engineer and AI researcher who has also taught and studied abroad in two foreign countries, he is passionate about the applications of technology to second language learning and assessment. His work in AI specializes in the field of Natural Language Processing (NLP), where he has made innovative contributions to automatic readability estimation, automatic writing evaluation, and estimating item response theory (IRT) item parameters for L2 assessments using explanatory models with NLP features.

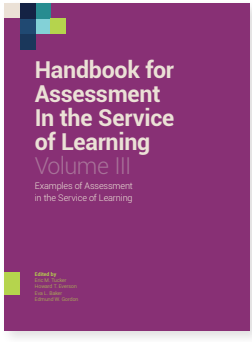
Jessica W. Younger, Ph.D., is an educational neuroscientist dedicated to developing effective interventions that empower learners to reach their full potential. With over a decade of experience, her work explores how individual differences shape learning, leveraging advanced statistical modeling and large-scale data analysis to personalize education. Currently, as Senior Manager of Research Products at PBS KIDS, Younger leads efforts to optimize educational content through innovative research tools, data-driven insights, and experimental platforms. Throughout her career, she has led multidisciplinary teams in designing research platforms, digital assessments, and large-scale studies that examine cognitive development and learning variability. Her work spans executive function, digital interventions, and personalized learning, with a focus on translating research into actionable insights for educators, technologists, and policymakers. By integrating neuroscience, data science, and education, Younger remains committed to advancing the understanding of how people learn best—ensuring that educational approaches are inclusive, evidence-based, and tailored to the needs of diverse learners.

Constance Yowell is senior advisor to the provost for special projects at Northeastern University. She previously served as senior vice chancellor for educational innovation, where she led the university's Center for Advancing Teaching and Learning Through Research, the University Honors Program, Undergraduate Research and Fellowships, Employer Engagement and Career Design, the Global Experience Office, Peer Tutoring, Self-Authored Integrated Learning, and the PreMed and PreHealth Advising Program. Before joining Northeastern, Yowell served as executive vice president of Southern New Hampshire University where she oversaw community engagement and outreach, with a focus on engineering a stackable, personalized learning approach for low-income, first-generation learners. Yowell began her career as an associate professor at the University of Illinois after serving as a policy analyst in the New York City school system and the U.S. Department of Education. Her research and policy work have focused on the deep disparities in local and federal education systems, particularly for African American and Latinx students, and she has written prolifically on the impact of educational policies and equity on student outcomes. Yowell holds a Ph.D. in child and adolescent development from Stanford University and a bachelor's degree from Yale University.

Handbook for Assessment in the Service of Learning Series







Volume III of the *Handbook for Assessment in the Service of Learning* bridges the gap between aspiration and application, by translating core design principles into practice through a collection of examples. This volume presents tangible "existence proofs" from a broad range of educational contexts—including digital learning platforms, PreK–12 classrooms, game-based learning environments, and skills-based credentialing programs. Each worked example can be understood through three complementary lenses: assessment as an evidentiary argument, as a feedback loop, and as a social practice. This framework reveals how thoughtfully designed assessment systems with actionable feedback can balance the need for evidence of learning. By showcasing assessments that are seamlessly integrated with learning and instruction, this volume advances the proposition that to assess is, fundamentally, to teach and to learn. It offers practical models and designs that embed assessment within instruction to cultivate skills and support meaningful learning.