Handbook for Assessment in the Service of Learning

Volume II

Reconceptualizing Assessment to Improve Learning

Edited by
Stephen G. Sireci
Eric M. Tucker
Edmund W. Gordon

Handbook for Assessment in the Service of Learning

Volume II

Reconceptualizing Assessment to Improve Learning

UMassAmherst
University Libraries

Edited by
Stephen G. Sireci
Eric M. Tucker
Edmund W. Gordon

Series Editors

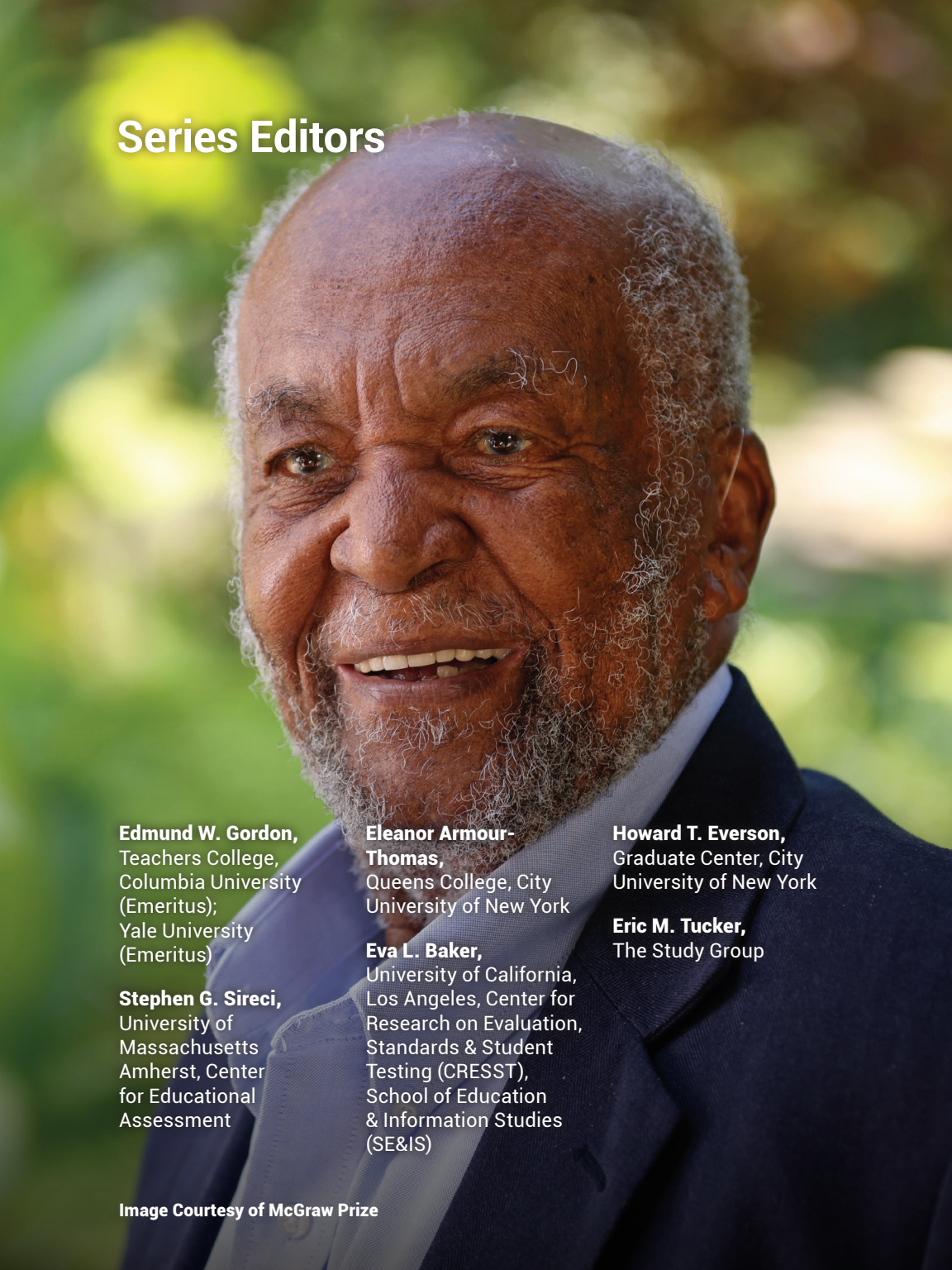


Image Courtesy of McGraw Prize

UMassAmherst

University Libraries

Handbook for Assessment in the Service of Learning, Volume II: Reconceptualizing Assessment to Improve Learning ©

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

DOI: 10.7275/ejm6-se46
ISBN: 978-1-945764-34-9

Cover Design by Dezudio
Book Design by The Study Group

The suggested citation for this handbook is: Sireci, S. G., Tucker, E. M., & Gordon, E. W. (Eds.). (2025). *Handbook for Assessment in the Service of Learning, Volume II: Reconceptualizing Assessment to Improve Learning*. University of Massachusetts Amherst Libraries.

Contents

Volume Contributors

Sergio Araneda, University of Massachusetts Amherst

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)

Héfer Bembenutty, Queens College, City University of New York

Randy E. Bennett, ETS, Research Institute

Mary K. Boudreaux, Southern Connecticut State University

Carol Bonilla Bowman, Ramapo College of New Jersey

Susan M. Brookhart, Duquesne University

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)

Danielle Crabtree, University of Massachusetts Amherst

Kerrie A. Douglas, Purdue University

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

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

James Paul Gee, Arizona State University (Emeritus)

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

Norris M. Haynes, Southern Connecticut State University

Neal Kingston, University of Kansas

Kimberly McIntee, University of Massachusetts Amherst

Maria Elena Oliveri, Purdue University

Mya Poe, Northeastern University

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)

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

Eric M. Tucker, The Study Group

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, Dr. Pamela Cantor M.D., 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, Bill 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, 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 60 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 re-purpose 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 they mean 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 mispreparing 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. https://www.ets.org/Media/Research/pdf/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., 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, 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 US.

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.

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*, 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

Reconceptualizing Assessment to Improve Learning

Eric M. Tucker and Stephen G. Sireci

"Never forget the world of the possible." I wrote down this quote from my friend and mentor, Professor Edmund W. Gordon, on February 23, 2024. It was about three years after he first invited me to work with him on a "Handbook for Assessment in the Service of Learning." Of course, I agreed to co-edit the Handbook—little did I envision just how far that initial idea would expand. What began as a plan for a single handbook has blossomed into a full series of Handbooks on Assessment in the Service of Learning. I am proud to have joined an all-star editorial team in bringing forth Volume II of this series. This journey has shown me *the world of the possible* in educational assessment, a world I would not have imagined just four years ago.

Stephen G. Sireci
Northampton, MA

We are proud to offer Volume II of the *Handbook for Assessment in the Service of Learning*, to all those who strive to help others through education. This volume, entitled *Reconceptualizing Assessment to Improve Learning*, serves a special role in this series. Volume I of the Handbook explored the foundational design principles and research bases for transforming assessment to inform teaching and learning processes, essentially making the case for why change is needed and outlining key design imperatives. Volume III, at the other end of the arc, will illustrate practical implementations and working examples–case exemplars of assessment approaches that embody aspects of these new approaches. Volume II stands as the conceptual and methodological bridge between these two. In these pages, we move from Volume I's focus on research, design, and technology to reconceptualization and innovation, redefining what assessment can be and providing prototypes of how to do it. Our focus here is on abandoning outdated traditions of educational testing in favor of approaches to assessment that serve teachers and learners first. The chapters assembled in Volume II–contributed by an all-star cast of authors at the cutting edge of the field–offer examples of how we can design assessments that *truly* support learning: how we can harness new technologies to improve assessment, ensure our assessments meet the needs of all learners, and provide richer information for students, educators, and other stakeholders invested in education. In short, this volume tackles *how* we might reconceptualize assessment to fulfill the ambitious vision articulated by Professor Edmund W. Gordon and the Gordon Commission over a decade ago. The Gordon Commission on the Future of Assessment in Education (2013) argued that traditional testing–fixated on ranking students and certifying "what is"–must give way to approaches that illuminate how learning happens and how it can be improved. Our task in Volume II is to build on that vision, providing aspects of both the conceptual blueprints and the inventive tools needed to reinvent assessment in the service of learning.

To organize this rich body of work, Volume II is divided into two sections. Section I: Foundations and Frameworks for Learner-Centered Assessment lays out the key theories, principles, and frameworks guiding the transformation of assessment. These chapters articulate why we must reconceptualize assessment and on what bases–from formative assessment foundations to considerations of validity and social justice. Section II: Innovations in Practice–Tools and Methods Serving Learning then showcases a variety of cutting-edge approaches that put those

principles into action. The chapters in Section II present novel methodologies and tools–from game-based assessments and learner-centered portfolios to culturally responsive co-design and new uses of data–each illustrating how assessment can be embedded into educational practice to engage learners and provide meaningful feedback for improvement. In essence, Section I gives us the "why" and "what" of reconceptualizing assessment, while Section II explores some of the practical "how"–mirroring our series' progression from Volume I's foundational research and design approaches to Volume III's applied cases.

Section I: Foundations and Frameworks for Learner-Centered Assessment

Section I curates aspects of the conceptual underpinnings for "assessment in the service of learning." The six chapters in this section establish core approaches and big-picture ideas that set the stage for reimagining assessment as a tool to inform and improve learning, rather than merely to audit what Professor Gordon might call achieved intellective competence. These chapters span formative assessment, self-regulated learning, personalization and equity, and theoretical frameworks for validity and justice–and together provide a meaningful contribution to a foundation for a learner-centered assessment approach.

- **Susan M. Brookhart:** ***Developing Educational Assessments to Serve Learners:*** Susan Brookhart (2025) provides a perfect beginning for this volume. She reminds us that learning begins long *before* children enter formal schooling, and that formative assessment is the solid foundation on which all assessments in the service of learning are built. Brookhart identifies key factors needed to facilitate assessment for learning–a supportive learning culture, clear learning goals, and clear success criteria–underscoring what must be in place for an assessment to successfully serve learners. She follows these insights with practical guidance on creating assessments that yield the feedback students need to advance their learning. Her chapter reinforces the notion that a formative, feedback-rich culture is foundational to any effective learner-centered assessment system.
- **Héfer Bembenutty:** ***Toward a Culturally Self-Regulated Dynamic Pedagogy Assessment System*****:** In the next chapter, Héfer Bembenutty (2025) continues the theme of integrating curriculum, instruction, and assessment to support self-regulated learning. Extending Armour-Thomas and Gordon's dynamic pedagogy framework, Bembenutty describes an approach that values students'

cultures and empowers learners to use assessment information to understand and guide their own learning (a model he refers to as *Culturally Self-Regulated Dynamic Pedagogy*). The assessment-pedagogy practices outlined in this chapter go beyond simply adding formative assessments into instruction—they emphasize creating inclusive learning environments where assessment is woven into the learning process and students are active agents in their learning. By demonstrating how teaching, learning, and assessment processes can jointly foster students' self-regulation skills, this chapter exemplifies the deep integration of assessment with instruction to benefit learners.

- **Randy E. Bennett, Eva L. Baker, and Edmund W. Gordon: *Personalizing Assessment for the Advancement of Equity and Learning:*** Bennett, Baker, and Gordon (2025) also highlight culturally responsive principles and the importance of learner variation. This chapter illustrates how assessment in the service of learning can be designed to advance equity by personalizing the assessment process. The authors propose using personalized assessments to accommodate the wide range of variation in the learner population. They review research on learner variability and describe different conceptualizations of diversity, then offer concrete principles for adapting both learning activities and assessments to build on learners' individual experiences, cultures, and identities. These principles form a helpful roadmap for designing assessments that can flexibly meet the needs of diverse learners, ensuring that assessment practices contribute to equity and do not treat fairness and responsiveness as an afterthought.
- **Norris M. Haynes, Mary K. Boudreaux, and Edmund W. Gordon: *A Theory-Informed and Student-Centered Framework for Comprehensive Educational Assessment:*** Haynes, Boudreaux, and Gordon (2025) present a broad theoretical framework to guide learner-centered assessment. They draw on three major perspectives—constructivism, sociocultural theory, and implicit theory—to ensure that assessments for learning provide valid insights into *how* students learn and develop. By acknowledging the influence of school and classroom culture and climate (including the constraints of the "hidden curriculum"), this chapter shows how assessments can make more meaningful learning experiences that support the holistic development of all learners. The authors discuss a comprehensive range of assessment types (formative, summative, diagnostic, ipsative, self-assessment, norm- and criterion-referenced, curriculum-based, etc.), illustrating how each can be employed,

in line with sound learning theory, to support student growth. This wealth of approaches, grounded in the learning sciences, enriches our toolbox for designing assessments that are both rigorous and learner-centered.

- **Stephen G. Sireci and Danielle M. Crabtree: *Validity Theory and Validation of Assessments in the Service of Learning:*** In this chapter, Sireci and Crabtree (2025) tackle questions of validity for assessments whose primary purpose is to serve learning. They explain traditional notions of test validity and how classic validity theory applies when assessments are repurposed to support learning. They discuss how to gather and evaluate validity evidence to ensure assessments in the service of learning are actually accomplishing their intended goals. The chapter aims to rectify the glaring lack of practitioner (teacher) perspectives missing from many test development and validation processes. Their chapter draws largely on the established standards of educational testing (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), but also introduces newer perspectives tailored to the specific goals of using assessment to enhance learning for students. Ultimately, they argue that transforming assessment does not mean abandoning rigor; rather, it requires expanding our concept of rigor to include usefulness for learning. Assessments in the service of learning must meet high standards of technical quality *and* yield information that is instructionally actionable and meaningful to students and teachers.
- **Stephen G. Sireci, Sergio D. Araneda, and Kimberly A. McIntee: *Social Justice in Educational Assessment:*** A Blueprint for the Future**:** Sireci, Araneda, and McIntee (2025) round out Section I with a chapter that poses a provocative, forward-looking question: *How can assessment be reimagined as a tool for social justice in education?* In this concluding chapter of Section I, they argue that issues of fairness, equity, and justice should not be afterthoughts in assessment design—they must be treated as foundational principles from the very start. They examine ways in which current assessment practices can unintentionally perpetuate inequities (for instance, through cultural bias in test content, unequal access to test preparation, or high-stakes uses that disproportionately impact marginalized groups). They also outline strategies to ensure assessment systems promote equity and empowerment rather than reinforce disparities. These strategies include designing culturally responsive

assessments that value diverse ways of knowing, involving students and communities in co-designing assessments, and using assessment data proactively to identify and close opportunity gaps (instead of using data punitively to label or punish). In essence, this chapter elevates the conversation to the policy and ethical level, extending the learner-centered assessment narrative into the realm of social responsibility. They contend that a truly learning-centric assessment system must also be a justice-centric system–one that actively works to dismantle historical biases and create more inclusive, supportive educational environments. By articulating concrete principles and recommendations for socially just assessment, they provide both a moral compass and a practical guide for the future. This chapter challenges us to ensure the transformation of assessment remains aligned with the broader goal of educational equity.

Although each of these chapters has a distinct focus, together they offer components of an emerging framework for rethinking assessment. They collectively prompt us to reconsider *why* we assess (our purposes), *what* we assess (the constructs and competencies we value), and *how* assessment practices impact learners and society (the consequences we care about). In doing so, Section I lays robust groundwork for reinventing assessment as a positive force in the learning process. Volume I gave us the vision and the design imperatives for assessment reform, and here in Volume II–especially through Section I–we undertake the critical work of reconceptualization, redefining the fundamental ideas and frameworks on which future assessments will be built. These conceptual foundations now pave the way for the innovations presented in Section II, where theory meets practice.

Section II: Innovations in Practice–Tools and Methods Serving Learning

If Section I explains why and on what insights we must change assessment, Section II explores important aspects of how those principles can be realized through new approaches and tools. The chapters in Section II showcase a range of innovative practices that embed assessment into the fabric of teaching and learning. From games and portfolios to data analytics and co-designed assessments, each contribution breathes life into the learner-centered vision by demonstrating concrete strategies for making assessment an integral, engaging part of education. Collectively, these chapters show that the lofty ideals outlined in Section I can indeed be translated into inventive designs that improve learning.

- **James Paul Gee: *Game-Based Learning: A Design-Based Theory of Teaching–Learning–Assessment Systems:*** Jim Gee (2025) opens Section II with a fascinating illustration of how assessment can be seamlessly integrated with instruction to foster engagement and deep learning. He presents a design-based theory of "teaching–learning–assessment" systems grounded in what we can learn from good games. As Gee insightfully observes, "good games are good for teaching, learning, and assessment". In a well-designed game, a player's constant problem-solving and immediate feedback naturally generate evidence of learning; the assessment is essentially woven into the gameplay itself. Gee's chapter explains how games can integrate teaching, learning, and assessment invisibly (to the learner) yet effectively, and he provides a blueprint for developing such game-based assessment systems. Beyond theory, this chapter offers practical design principles for educators and developers interested in creating engaging, game-like assessments that motivate learners and simultaneously yield rich information about their learning processes.
- **Carol A. Bowman and Edmund W. Gordon: *The Educative/Learning Portfolio:*** Towards Educative Assessment in the Service of Human Learning**:** Bowman and Gordon (2025) reintroduce a more familiar, yet underutilized, assessment tool–the portfolio–and reconceptualize it as an "educative portfolio." They describe how a student's portfolio of work can be transformed from a static showcase of accomplishments into a dynamic process and instructional tool that actively cultivates learning. In an *educative portfolio* model, compiling and reflecting on one's work becomes an integral part of the learning experience itself. Students select pieces, reflect on their growth, and discuss their work, meaning the assessment happens through those activities. This approach has the potential to yield rich, authentic evidence of learning–in fact, the portfolio artifacts provide "more useful and abundant evidence of achievement than a simple metric," offering a revealing window into the processes of the student's learning. Unlike the "stealth" assessment in a game, portfolio-based assessment is purposefully visible and transparent: clear objectives, expectations, and reflective actions before, during, and after the assessment are central. Bowman and Gordon show how transparency and reflection in portfolio assessment support learning, making the process educative for the student. Like Gee's games, the portfolio chapter exemplifies integrating assessment with curriculum and instruction–albeit in a different form–and demonstrates that even traditional assessment formats can be innovated to serve learning more effectively.

- **Maria Elena Oliveri, Kerrie A. Douglas, and Mya Poe: *Building Culturally and Linguistically Responsive Workplace Assessments for Learning:*** Oliveri, Douglas, and Poe (2025) advance the idea of culturally responsive assessment through the lens of workplace learning. They illustrate how involving learners (and other stakeholders) directly in the test development process–for instance, via participatory co-design–leads to assessments that are more valid and appropriate for diverse populations. After introducing the concept of culturally and linguistically responsive teaching and assessment, the authors present a compelling use case from engineering education. In this example, assessments were co-designed to reflect multilingual, multicultural workplace realities. The chapter demonstrates that when assessments are grounded in learners' cultural contexts and allow multiple ways for learners to demonstrate competence, the assessments become not only fairer but also more instructionally valuable. In the context of workplace learning, this means assessments better prepare and reflect what learners need on the job, while honoring the diverse backgrounds they bring. Oliveri, Douglas, and Poe offer practical guidance for developing assessments *with* learners rather than *for* learners, embodying the principle that assessment design should adapt to learners (instead of expecting learners to adapt to rigid assessments). This culturally responsive co-design approach demonstrates how we can develop assessments that genuinely include and empower every learner.
- **Stephen G. Sireci and Neal Kingston: *Removing the "Psycho" from Education Metrics:*** In this provocatively titled chapter, Sireci and Kingston (2025) examine aspects of how assessment results are reported and used. They critique traditional testing metrics and reporting formats, which too often mystify or alienate educators and students by drowning them in psychometric complexity. Their chapter advocates for reimagining how assessment results are communicated to serve learning, and they argue for shifting from obscure statistics to intuitive, learner-centered feedback that students, teachers, and parents can readily understand and act upon. As a working example, they describe the *Dynamic Learning Maps* system–an innovative assessment designed for students with significant cognitive disabilities–as an illustration of assessment geared toward diagnosing individual learning needs and guiding instruction, rather than merely cataloging deficits. This system aims to provide rich diagnostic profiles of what a student *can do* and what might help them progress next, exemplifying assessment as a supportive tool for

learning. Throughout the chapter, they show how tests and score reports can be designed in plain, user-friendly language without sacrificing the depth of information. By redesigning score reports to emphasize clear, actionable insights (what skills a student has mastered, and what they should work on next), they illustrate how to maintain rigor while making data more useful. The message across this chapter (and Oliveri et al.'s as well) is that assessments can and should adapt to learners and educators—not the other way around—by providing information that is accessible, meaningful, and geared toward helping every student learn.

- **Gregory K. W. K. Chung, Tianying Feng, and Elizabeth J. K. H. Redman: *Using Learner-System Interactions as Evidence of Student Learning and Performance:*** Chung, Feng, and Redman (2025) push the frontier of assessment by asking: What if *every* interaction a learner has with educational materials could count as assessment data? In this final chapter of Section II, they explore how emerging technologies and data analytics enable entirely new forms of evidence of student learning. The authors posit that every click, response, or choice a student makes in a digital learning environment is potential data about their thinking and skills. By capturing these fine-grained learner–system interactions, for example, how a child approaches problems in an online math game, we can glean insights that no traditional test alone could offer. Chung et al. outline methods for identifying which learner behaviors to capture, how to record them, and how to analyze this flood of data to draw valid inferences about student learning. Their approach applies rigorous measurement principles to forms of behavioral data not typically considered in assessment. They show how students' interactions with digital tutoring systems, educational games, and other instructional software can be interpreted as assessment evidence and modeled to provide ongoing feedback. The chapter includes vivid examples, such as observing preschoolers' strategies in a math game, to illustrate how these interaction data can reveal learning processes and guide instruction in real time. Although the context of their example is early childhood mathematics, the underlying principles generalize to all levels and subjects: modern technology allows us to embed assessment into virtually any learning activity. This work highlights how assessment is evolving into something much broader than tests—it can encompass the *continuous stream of data* generated by learners as they engage with learning materials.

Each of these chapters advances understanding of practical approaches to make assessment more integrated with learning. Taken together, the contributions in Section II span a remarkable range of contexts and methods—but they all show how the core ideals from Section I can be realized in practice. Whether through immersive games, reflective portfolios, co-designed culturally responsive tasks, reimagined score reports, or data-rich digital environments, these authors are breathing life into the idea of assessment as a tool for learning. They exemplify the creativity and dedication needed to turn assessment from a once-a-year audit into an ongoing, student-centered conversation about growth.

Emergent Themes Across Volume II

Across both sections of Volume II, several key themes reverberate, weaving a unifying narrative of what it means to make assessment truly learner-centered:

- Formative feedback and improvement: A shift toward feedback-rich, formative practices (Shute, 2007) is evident throughout the volume. From Brookhart's emphasis on a formative culture to Bembenutty's focus on real-time self-regulation, the idea of using assessment to provide continuous feedback for improvement is a common thread. Even in Section II's tools, we see this theme: Gee's game-based assessments offer frequent feedback in context, Bowman and Gordon's portfolios embed feedback through reflection, and Chung et al.'s analytics turn interactions into actionable feedback. The notion that assessment should *inform* and *guide* learning—rather than merely judge it—underpins these chapters. (Gordon Commission on the Future of Assessment in Education, 2013; Gordon, 2020)
- Learner agency and engagement: The authors in this volume consistently treat learners as active participants in the assessment process, not passive subjects of measurement. This commitment to *learner agency* shows up in many forms: Bembenutty's culturally self-regulated pedagogy empowers students to monitor their learning, Oliveri et al.'s participatory co-design actively involves learners in creating assessments, and portfolio assessment (Bowman & Gordon, 2025) gives students voice and choice in showcasing their learning. Even Gee's games put the learner in charge of problem-solving within the assessment environment. In line with Professor Gordon's vision (Cauce & Gordon, 2013), the student is not a mere object of assessment but an agent whose engagement, self-management, and self-reflection are integral to the

process. By fostering agency, these chapters suggest that assessment can actually *motivate* and empower learners.

- Validity, fairness, and social justice: A strong imperative around quality measures and the advancement of justice runs through Volume II. Nearly every chapter grapples with how to make assessment more fair, inclusive, and beneficial for *all* learners. Bennett et al. explicitly center diversity by personalizing assessment to learner needs; Haynes et al. emphasize culturally attuned frameworks and holistic validity; Oliveri et al. design for linguistic and cultural responsiveness in diverse contexts; and our social justice chapter insists on making fairness and justice fundamental criteria for any assessment system. At the same time, maintaining validity and rigor is a shared concern–the volume does not advocate diminishing the importance of technical quality, but rather expanding our definitions of quality. Sireci and Crabtree's chapter, for example, shows how we can uphold rigorous validation standards for new kinds of assessments, ensuring that innovative assessments yield trustworthy evidence about student learning *while* avoiding cultural bias or misuse. In sum, Volume II envisions assessment systems that are high quality, rigorous, and just–assessments that earn stakeholders' confidence through validity *and* demonstrate a commitment to fairness and social responsibility.
- Dynamic integration of assessment, learning, and instruction: A recurring theme is the blurring of lines between assessment and instruction. Many authors echo Professor Gordon's call to integrate assessment in the processes of instruction rather than treat it as a separate, after-the-fact event. (Armour-Thomas & Gordon, 2013, 2025) Brookhart and Bembenutty set the stage by describing classroom cultures where assessment is part of everyday teaching and learning. In Section II, this integration becomes concrete: in Gee's chapter, assessment *is* the gameplay; in Bowman's, assessment is woven into the act of curating and reflecting on learner work; in Chung et al's, assessment data is captured as students learn in digital environments. The benefit of such integration is twofold: it makes assessment more natural and less anxiety-provoking, and it aims to produce more instructionally relevant data. Throughout the volume, we see that integrating assessment with instruction has the potential to lead to more timely insights and create a more supportive experience for learners–fulfilling the ideal of assessment as a "pedagogical transaction" embedded in learning.

- Responsiveness and relevance: Finally, responsiveness and the cultural foundations of learning emerge as a vital theme (Bennett et al., 2024; Mislevy et al., 2024; Nasir et al., 2020). The chapters collectively recognize that learners bring significant variation in backgrounds, languages, and ways of knowing to the table, and that assessments must honor and reflect that variation. This is most explicit in works like Oliveri et al.'s co-designed assessments for multicultural settings and Bennett et al.'s personalized approaches for diverse students. Haynes et al. also incorporate sociocultural perspectives to ensure assessments are meaningful across different contexts, and our social justice chapter takes this further to address systemic biases. Even outside the equity-focused chapters, cultural relevance appears in Gee's attention to engaging all learners through game narratives and in portfolio assessment's accommodation of individual expression. The through-line is that one-size-fits-all assessments are no longer acceptable; to truly serve learning, assessment practices must be adaptable to cultural and individual variation. By making assessments more responsive to learners' contexts, we not only improve fairness but also make assessment results more meaningful and actionable for each learner.

These themes–formative feedback, learner agency, validity and fairness, dynamic integration with instruction, and cultural responsiveness–resonate throughout portions of Volume II and tie the chapters together. They reflect a shared commitment to redefining assessment as something fundamentally in service of learning and human development.

Looking Ahead to Volume III

As we begin Volume II, it is worth reflecting on how the insights gathered here might set the stage for the third volume in our series. Volume III will carry this work forward by showcasing implementations and exemplars of assessment in the service of learning. In Volume III, readers will see the concepts and innovations from Volumes I and II come alive in various teaching and learning contexts. We will explore case studies and models of assessment systems that have been implemented, demonstrating the impact and feasibility of the ideas we've been discussing. In a sense, if Volume II provides frameworks and tools, Volume III will show some functioning working examples–works in progress–built upon those designs principles.

The chapters of Volume II provide key tools in a conceptual and methodological toolkit for readers seeking to transform assessment. They have illustrated aspects of "the world of the possible" that was beyond our own vision when this project began. Now, Volume III will challenge us to apply that toolkit and learn from concrete experiences. By connecting theory to practice, Volume III will complete the bridge that Volume II has built between foundational principles and practical realization. We are excited to see how innovative assessments *in action* can further validate these ideas, reveal new challenges, and inspire continued refinement of assessment for learning.

In closing, we feel a deep sense of gratitude. First, we want to thank the authors of these chapters, who are pioneers in our field–their dedication and creativity made this volume possible. We are grateful as well to our fellow editors and collaborators; working with an editorial team of such vision and expertise has been a privilege. Above all, we thank Professor Edmund W. Gordon—the Professor—whose unwavering vision has guided this journey from the start. Over twelve years ago, Professor Gordon wrote of the coming "conflict and contradiction" between traditional assessment practices and new scientific developments. He challenged us to resolve that conflict by reimagining assessment in bold ways. This Handbook series is, in many ways, one response to that challenge. As Gordon and the Gordon Commission foresaw, science, technology, and imagination have opened new possibilities for assessment. The work in Volume II represents one collective effort to advance understanding of how we might realize those possibilities–to redesign assessment in light of what is now *possible* for the betterment of learners.

With profound gratitude and optimism, we invite you to engage with the chapters of this volume. We hope you find in them not only rigorous scholarship and practical insight, but also the same sense of hope and inspiration that we have found. Together, may we continue to explore and expand *the world of the possible* in assessment, in service of every learner's growth.

References

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

Armour-Thomas, E., & Gordon, E. W. (2013). *Toward an understanding of assessment as a dynamic component of pedagogy*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/armour_thomas_gordon_understanding_assessment.pdf

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.

Bembenutty, H. (2025). Toward a culturally self-regulated dynamic pedagogy assessment system. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Bennett, R. E., Baker, E. L., & Gordon, E. W. (2025). Personalizing assessment for the advancement of equity and learning. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (Eds.). (2024). *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Routledge.

Bowman, C. A., & Gordon, E. W. (2025). The educative/learning portfolio: Towards educative assessment in the service of human learning. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Brookhart, S. M. (2025). Developing educational assessments to serve learners. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Cauce, A. M., & Gordon, E. W. (2013). *Toward the measurement of human agency and the disposition to express it*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/cauce_gordon_measurement_human_agency.pdf

Chung, G. K. W. K., Feng, T., & Redman, E. J. K. H. (2025). Using learner-system interactions as evidence of student learning and performance. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Gee, J. P. (2025). Game-based learning: A design-based theory of teaching–learning–assessment systems. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

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.

Haynes, N. M., Boudreaux, M. F., & Gordon, E. W. (2025). A theory-informed and student-centered framework for comprehensive educational assessment. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Mislevy, R. J., Oliveri, M. E., Slomp, D., Crop Eared Wolf, A., & Elliot, N. (2024). 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. 151–167). Routledge.

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

Oliveri, M. E., Douglas, K., & Poe, M. (2025). Building culturally and linguistically responsive workplace assessments for learning. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Sireci, S. G., Araneda, S. D., & McIntee, K. A. (2025). Social justice in educational assessment: A blueprint for the future. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Sireci, S. G., & Crabtree, D. M. (2025). Validity theory and validation of assessments in the service of learning. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Sireci, S. G., & Kingston, N. (2025). Removing the "psycho" from education metrics. In S. Sireci, E. Tucker, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning: Volume II*. University of Massachusetts Amherst Libraries.

Shute, V. J. (2007). *Focus on formative feedback*. Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2007/hslv.html

VOLUME II | SECTION 1

Foundations and Frameworks for Learner-Centered Assessment

Developing Educational Assessments to Serve Learners

Susan M. Brookhart

Abstract

In this chapter I examine aspects of educational assessment that could or should change or evolve to serve the needs of learners more effectively in the future. I approach this task in three main sections: by considering (1) what learners need; (2) how assessment can help meet these needs (i.e., examining how assessment meets learners' needs now and how that could be enhanced by new developments in educational tools and processes); and (3) how validity and validation concerns might develop accordingly. A final section summarizes these points in relation to three of the seven Principles for Assessment in the Service of Learning and makes recommendations to inform the development of assessment tools and processes that better support learners in their learning.

Children arrive at school as young—sometimes very young—people who have already accomplished a great deal of learning. Much of that learning was developmental and learned in the context of a family or other care group: learning to walk, learning to speak, learning to whom to turn for food and emotional support. Children add an additional learning context to their lives when they come to school. Cognitive outcomes become more salient in school (Resnick, 1987). As students age, learning becomes less physical and visual, and more mediated by language, but students still need to be healthy, safe, engaged, supported, and challenged (Slade & Griffith, 2013). They still need a learning community. Shepard et al. (2018) argued that sociocultural theory acknowledges this need by positing that self-regulation, self-efficacy, sense of belonging, and identity are interwoven with cognitive development. School learning is situated in classroom learning communities, somewhat similar to the way preschool learning was situated in a family or care group.

The primary assessment from a learning perspective is classroom formative assessment; other assessments for other purposes (e.g., accountability assessment) need to be coherent with and connected to formative assessment (Shepard et al., 2018). Formative assessment is assessment that occurs during learning, providing information to students and teachers that can move the learning forward. Formative assessment involves both a process and some sort of instrument, tool, or method (Bennett, 2011). The formative assessment process used in a classroom learning community will best meet students' learning needs if it uses the formative learning cycle (Where am I going?—Where am I now?—Where to next?) so that students can activate cognitive, affective, and behavioral regulation strategies to move them toward their learning goal (Andrade & Brookhart, 2019; Andrade et al., 2021), developing their evaluative judgment (Panadero et al., 2019).

To adopt a sociocultural stance and recognize that learning is done by whole persons does not negate all prior theories—for example, it is still useful to study and understand students' cognitive structures (Shepard et al., 2018; Wiliam, Fisher, & Frey, 2024). What a sociocultural framing does is allow us to focus on students' self- and co-regulation of learning as primary aspects of student learning needs and on features of formative assessment and feedback as primary ways in which assessment meets those needs (Bailey & Heritage, 2018). These aspects are detailed in the next two sections.

What Do Learners Need in Order to Learn?

Learners need a supportive classroom learning culture (sometimes called the classroom learning environment; Ames, 1992). They also need clear learning goals and success criteria, shared with them in such a way that they can activate self- and co-regulatory processes to actively move themselves closer to the learning goal. In classrooms, this is often called the formative learning cycle (Brookhart & DePascale, in press; Brookhart & McTighe, 2017). Finally, they need high-quality feedback from teachers, self, peers, and sometimes from computers or other learning materials and, importantly, the opportunity to use that feedback. Effective feedback and its use are part of the formative learning cycle, but feedback is so important that it warrants its own discussion, which appears in a subsequent section of this chapter.

Supportive Classroom Learning Culture

Learners need a classroom culture that views mistakes as opportunities to learn and encourages learning together. The idea that classroom culture can contribute to students' motivation to learn has been around a long time and can be traced to theories about students' learning goal orientations (Ames, 1992). How assessment is used is an important element in determining whether students perceive their classroom as a learning culture or an evaluative culture (Ames, 1992; Crooks, 1988).

Students need to think of themselves as competent learners who belong in a community of others who are also active learners. An important feature of a learning community is how students view being wrong and the role of productive struggle in learning (McMillan, 2018). Leighton et al. (2013) proposed one way to create a safe learning environment is to explain to students how making and then understanding errors has value for learning. This explanation helps students expect to make errors and makes it safe for them to discuss their errors as they learn complex content. While some learning can be accomplished individually, for example memorizing math facts, deeper and more meaningful learning is best accomplished in a community where students can work with peers, including more and less experienced peers, on learning tasks (Laal & Ghodsi, 2012).

To benefit from instruction, learners need appropriate background knowledge and experience. They need to know their own cultural background knowledge and experience is honored. Giving students voice and choice in assessment is

an important way to bring students' cultural background into instruction and assessment (Ladson-Billings, 2014; Taylor & Nolen, 2022). Ladson-Billings (1995) used the term "culturally relevant pedagogy" to describe teaching that builds on connections between teachers and students' families, communities, and daily lives. More recently, others have added to the insights that students need school—and assessment—to reflect their own cultures and funds of identity (Esteban-Guitart & Moll, 2014; Randall, 2021). This is an important part of creating a classroom culture of learning where students feel safe, welcomed, and supported in their learning.

The principle of active student involvement in learning in a classroom learning culture, and the underlying beliefs that students construct their own understandings, can pose a difficulty for teachers who are used to thinking "What am I going to teach?" instead of "What will students try to learn?" Studies of formative assessment in both pre-service and in-service teacher education have found that at both levels, programs in which teachers developed effective formative assessment practices were those that helped teachers shift their beliefs about student learning to realize that the assessment information needed to inform students' regulation of learning more than teachers' lesson plans (Brookhart, 2017).

Clear Learning Goals and Success Criteria

Given a supportive classroom learning culture, the first thing students need is to understand what learning goal they are pursuing (Chen et al., 2017; Heritage & Wylie, 2020). Having a goal is what makes the difference between student compliance—students simply doing what the teacher asks—and students' regulation of their learning. Regulation of learning requires having a learning goal (Zimmerman & Schunk, 2011). The goal needs to be specific enough that students have a clear sense of what they are trying to learn. This is typically accomplished by sharing expectations or criteria for what counts as evidence of learning (sometimes called success criteria), by sharing concrete examples of different levels of work with students, and by discussions and activities around the criteria and examples (Chen et al., 2017; James et al., 2006; Heritage & Wylie, 2020). During learning, students apply the learning goal and success criteria to engage in the self-regulation of learning (Moss, 2022).

The Formative Learning Cycle

Once a learning goal is set, learners need to set their sights on it, aim for it, and activate and sustain cognitions, affects, and behaviors in pursuit of the goal; in other words, they need to muster thoughts, motivation, and effort to regulate their learning (Zimmerman & Schunk, 2011). This metaphor of "aiming" is the image behind the term "learning target." It is not the goal, but rather the student aiming, that comprises regulation of learning. This regulation may be broader than self-regulation and include assessment information from teachers, peers, technology, and other learning materials (Andrade et al., 2021); such regulation is known as co-regulated learning and involves the whole classroom learning culture.

As a whole, this process is sometimes called the formative learning cycle (Brookhart & DePascale, in press; Brookhart & McTighe, 2017). The formative learning cycle is a practical instantiation of the process of the regulation of learning, inspired by three conditions of formative assessment originally proposed by Sadler (1989) and expanded into a three-question model of feedback by Hattie and Timperley (2007). In the formative learning cycle, students set a goal (Where am I going?); gather feedback on formative practice work from multiple sources to compare where they are and where they need to be (Where am I now?); and consider suggestions for next steps (Where to next?). Feedback targeted to a learning goal can lead to students increasing effort and motivation, seeking additional information, changing learning strategies, and restructuring cognitions (Hattie et al., 2021). Formative assessment is especially effective for learning when students initiate self-assessment (Lee et al., 2020), in other words, when they take ownership of the formative assessment cycle.

To regulate their learning, learners need to be engaged and active, paying attention, exerting appropriate effort, and employing metacognitive skills (Andrade et al., 2021). Giving students voice and choice supports these efforts (Taylor & Nolen, 2022). Hattie and Clarke (2019) described a feedback culture—what I have been calling in this chapter a classroom culture that supports learning—as one based on the formative learning cycle and characterized by students who have the "skill, will, and thrill" (pp. 12–13) to use feedback to move from surface learning to deep understanding, where they can relate a concept to other ideas and apply it in other contexts. That is, students need learning skill, for example knowing how and when to focus their work and thinking and in what direction, to move their work and understanding closer to the criteria. They need the will or disposition to exert the

effort needed to do this, based on the belief that this work will help them learn or make them smarter. They need the thrill or motivation to reach the success criteria, in other words, they need to be truly aiming toward the learning goal which has become their own and not just the teacher's.

Feedback and the Opportunity to Use It

Learners need descriptive, non-evaluative, ungraded feedback and opportunities to use the feedback to approach the learning goal (Brookhart, 2018; Hattie, 2009; Hattie & Clarke, 2019), differentiated according to the learner's proficiency level (Stobart, 2018), and meaningful to the learner (Taylor & Nolen, 2022). Several recent reviews of feedback research have shown that some types of feedback are more powerful than others. Outcome feedback, sometimes called knowledge of results or verification (Shute, 2008), is the simplest and most common type of feedback. Outcome feedback tells students whether they were correct or incorrect, or what their score or grade was. This kind of feedback is useful for some purposes, especially for tasks involving recall. Outcome feedback about correctness coupled with knowledge of the correct response, as for example on the back of math fact flash cards or in some computer learning software, can be effective for memory tasks (Mason & Bruning, 2001).

In contrast, cognitive feedback or elaboration (Shute, 2008) contains information that students can use in their thinking. Cognitive feedback helps students interpret the task, interpret their response or response processes in light of criteria, set goals and monitor progress, address particular errors, and envision next steps; in other words, descriptive, elaborated feedback supports the processes of the formative learning cycle. This kind of feedback generally has more powerful effects on learning (Hattie & Timperley, 2007; Shute, 2008; Van der Kleij et al., 2015). Whether elaborated feedback can be given depends in part on the cognitive demands of the assessment task. Complex tasks typically provide more opportunity for elaborated feedback because student responses include more evidence of student thinking than student responses to recall-level tasks. The quality of the assessment question or task is critical in supporting feedback and thus supporting learning.

Feedback is moot unless students have an opportunity to use it. The opportunity needs to be built into the instructional sequence. Some teachers mistakenly believe that students will make mental notes of their feedback and use it "next time." However, learners need concrete opportunities to use feedback to move

learning forward while they are still in the process of moving toward the intended learning goal. Jonsson and Panadero (2018) examined research on students' use of feedback in higher education and described three aspects of context that influence whether and how students will use feedback. One aspect is whether assignments are given in stages in higher education, where there is opportunity for feedback and the possibility of improvement. Similarly, in K–12 education students need formative practice work, timely feedback, and opportunities for revision (Chen et al., 2017). The second aspect affecting students' use of feedback is whether they have been taught how to do that. The third is whether descriptive feedback comes with other evaluative measures, like a score or grade. When feedback is accompanied by grades, students often focus on the grade and do not engage with the feedback (Winstone et al., 2016).

Section Summary

The main theme in this section is that learners need each other, they need content, and they need instruction and assessment that allows them to activate regulatory processes to become active, engaged students. Students need a classroom community in which they can learn with others, in which they feel safe to pursue learning challenges even when they might be wrong, and in which their own background and experience is honored. In this classroom community, optimal learning occurs when students exercise self- and co-regulation of learning by aiming for a learning goal, receiving feedback on how they are doing and suggestions for next steps, and having the opportunity to muster their cognitive, affective, and behavioral skills to use the feedback. The formative learning cycle is a key process.

How do assessments meet learners' needs?

Arguably the most important assessment features to meet these learner needs are high-quality assessment and learning tasks, attention to the formative learning cycle in the structure and sequence of classroom learning, feedback and scores that produce useful information for the students, and where possible the use of learning progressions to design assessments and interpret results. The following sections describe how high-quality assessments meet students' needs now; each section ends by suggesting how these features could be enhanced by new developments in assessment.

Assessment Questions and Tasks

Assessment questions and tasks help students learn when they are educative, engaging, provocative of student thinking, and relevant to the learner's culture and experience. Tasks are educative when they are related to real disciplinary thinking and are rich enough to provide both students and teachers with feedback they can use to improve performance (Heritage & Wylie, 2020, Wiggins, 1998). Such tasks help students in their formative learning cycle by instantiating what it means to understand or be able to do the kind of thinking or work implied by the learning goal. They also aid the formative learning cycle by provoking student responses that give evidence of that thinking and allow feedback on student thinking, not just correctness.

Tasks should be a clear match with the learning goal(s) assessed (Brookhart & DePascale, in press; Heritage, 2013). It does no good to share a learning goal and success criteria with students and then use assessments that do not line up with those goals and criteria. When assessment tasks clearly embody the desired learning outcomes, assessment tasks are also learning tasks (Carless, 2015). One recent study in higher education found the quality of assessment tasks affected the quality of feedback and participation, these variables affected student empowerment, strategic learning and self-regulation, and all variables directly or indirectly affected students' learning transfer (Ibarra-Sáiz et al., 2021).

The concept of coherence (Wilson, 2004) is also relevant here. All assessments used in a school or district should interpret learning goals at different levels—from classroom lesson-sized goals through large-scale standards-based accountability goals—in the same way, so what is taught is what is assessed, and students can recognize that in their learning and performance, from classroom formative assessment through large-scale assessment. This does not mean classroom assessment tasks must be the same as large-scale assessment tasks. In fact, they typically will not be, as classroom assessment tasks usually reflect smaller chunks of learning than large-scale assessment tasks. Rather, it means the underlying construct—what students are trying to learn—must be coherent throughout an assessment system. For example, does understanding the water cycle mean being able to list or draw its steps or write hypotheses about water-related problems (or both)? If a teacher's formative assessment interprets a standard one way and other assessments in the system interpret a standard in another, coherence is lacking and learners are confused.

Furthermore, assessment questions and tasks should be relevant to the learners' culture and experience (Ladson-Billings, 2014; Randall, 2021) in several ways. The performances required of students in performance-based assessment should be responsive to students' cultural differences and connected with students' lives in some way (Hood, 1998; Solano-Flores & Nelson-Barber, 2001). Moreover, assessment questions and tasks should draw on students' linguistic repertoires, as language mediates students' participation in assessment (Bailey & Durán, 2020).

There is plenty of room for improvement in future assessment developments. High-quality tasks that are a direct match with intended learning goals and rich enough to elicit student thinking are very difficult to craft. An improvement in the quality of assessment questions and tasks would be a huge benefit to learners. Needed is ongoing professional development for educators who design classroom assessments and for assessment vendors who design external assessments, as well.

Formative Assessment Processes and Tools

Formative assessment supports learners' processes, motivation, attention, engagement, effort, metacognition, and self-regulation (Andrade et al., 2019; Heritage & Wylie, 2020). As shown in the previous section, educational assessments that serve learners are primarily formative (Brookhart & DePascale, in press; Shepard et al., 2018). "Formative" describes an assessment purpose—in this case informing learning for both students and teachers—and not a particular assessment instrument, since many assessments can be used either formatively (to inform learning) or summatively (to certify or report learning).

The instrument, tool, or method used for formative assessment will be most effective for student learning if the question or task is clearly matched to the learning goal; if the criteria are clear and provided in a form that students can use, with training and instruction on how to use them; and if the assessment, whether formal or informal, is deployed in a process that supports the formative learning cycle. For example, a teacher might pause in an instructional sequence and have students self-assess their work using criteria, then provide opportunity for revision during which students can improve their work and deepen their learning (Chen et al., 2017).

Rubrics are a common way to present criteria and performance level descriptions in a form that students can use for self- and peer assessment and that teachers can use for providing feedback and deciding on next instructional moves. Research in higher education has shown rubrics help make the criteria for good work explicit for students (Jonsson, 2014; Nordrum et al., 2013) and students use rubrics for this purpose (Andrade and Du, 2005; Garcia-Ros, 2011). Similar conclusions have been drawn from research on rubrics in basic education (Brookhart, 2024). Other tools that present criteria in forms students can use include checklists or other lists of criteria, models, and demonstrations. The key seems to be that the criteria are available in a usable form for students, not necessarily that they are rubrics, and that students have instruction and practice in how to use them (Chen et al., 2017; Panadero & Alonso-Tapia, 2013).

Importantly, these reviews and studies have shown that without instruction and guidance in how to use rubrics, and opportunities to do so, students may misunderstand or misuse them. In the terms I have been using in this chapter, the process and tools need to work together to help students regulate their learning.

An obvious future assessment development that would prove useful in improving the use of formative assessment processes and tools is enhanced professional development for both in-service teachers, preservice teachers, and teacher education faculty. There is a lot of rhetoric around formative assessment in schools and teacher education institutions, but programs that do this well are still rare (Brookhart, 2017). Additional developments in the design and use of classroom formative assessment tools and processes would be useful, too, especially focused on strategies where students are the agents of their own assessment (Lee et al., 2021). As the quality of classroom formative assessment information improves, it can be added to the kind of data reviews many schools already do, allowing more specific, equitable, and effective learning diagnoses and instructional remedies than such reviews currently support (Oláh et al., 2010). Safir and Dugan (2021) call this "street data."

Assessment Scores and Feedback

In the previous section, "Feedback and the Opportunity to Use It" was listed as one of the things learners need. Feedback is included in this section as well because it also is something assessment can provide to meet learners' needs. Quantitative scores and qualitative feedback both produce information that is descriptive of

current learning status, correlated with learning goals, informational for taking next steps in learning, and connected to the student and their work (Brookhart, 2018; Shute, 2008). Feedback can have a powerful effect on student learning. However, not all feedback is effective in every case (Shute, 2008). In addition, effective feedback can differ markedly depending on the subject matter and the age and level of the student (Smith & Lipnevich, 2018).

Recent reviews of the feedback literature find that in general, the most effective feedback is descriptive information that feeds into the formative learning cycle by helping students understand the current quality of their work and making suggestions for steps they can take to improve (Hattie & Timperley, 2007; Shute, 2008; Van der Kleij, Feskens, & Eggen, 2015). Attention to recent reviews is important because the feedback literature extends back farther in time than many other research literatures in education; early research using behaviorist theoretical frameworks gave way to more cognitive and then sociocognitive and sociocultural models (Lipnevich & Panadero, 2021). As the definition of feedback changed from meaning the simple knowledge of results (right/wrong) needed for behavior-based studies of feedback to including the descriptions and suggestions needed to help students navigate the formative learning cycle, studies began to show increased effectiveness for feedback (Brookhart, 2018).

Future developments in assessment, therefore, should concentrate on equipping educators to provide appropriate feedback—typically descriptive comments based on shared success criteria, but sometimes scores, depending on the learning goal—at moments in an instructional sequence when acting on that feedback would move learning forward. It may be possible to use artificial intelligence to assist in this task. Also, equipping educators to craft and share (or co-create with students) clear success criteria matched to the learning goal will be key to moving forward, because effective feedback is based on those criteria. Future development could also include adding to the repertoire of available self- and peer assessment strategies.

Work is also being done on computer-based cognitive tutoring, which includes feedback to students and also dynamic cognitive modeling to provide feedback to the cognitive tutor itself (Ritter et al., 2007). To serve learners well, feedback from externally-produced learning software deployed to individual students in the classroom will need to be mediated by teachers designing lessons in the context

of a classroom learning culture—as opposed to, for example, just having individual students sit in front of computers.

Regarding enhancements in feedback from large-scale assessment that might ultimately serve learners, including attention to formative uses of that feedback, much work is currently being done in developing results reports with multiple pieces of information (Zenisky et al., in press), sometimes suggestive of additional instructional materials (e.g., Smarter Balanced, n.d.). Work is also being done on assessments that use cognitive diagnostic models and report diagnostic classifications based on probabilistic data (Bradshaw & Levy, 2019). The educators who use these reports need professional development to understand them and use them well.

Learning Progressions

Learning progressions are descriptions of hypothesized, and often empirically tested, increasingly sophisticated student understanding that result from ordered steps of instruction in school subjects (Mosher, 2022). Learning progressions help teachers interpret student thinking and learning and engage students in richer, more equitable learning experiences (Alonzo & Elby, 2019; Shepard, 2018). Of course, there are not enough educational psychologists in the world to develop an empirically verifiable representation (Graf & van Rijn 2016) of how children learn and develop in every domain taught in school, but in domains where they are available learning progressions help make interpreting assessment information, providing feedback, and supporting next steps more precise.

Learning progressions are particularly helpful as teachers create or select assessment questions and tasks, give feedback on student work, and plan next steps in instruction. Learning progressions can help teachers understand what differences to expect in students' responses to classroom formative assessments (Confrey, 2019). If the grain size of the descriptions in the ordered steps or levels in a learning progression are small enough to support lesson-sized decisions about what kind of growth typically comes "next" in students' understanding of a concept, assessment can support those decisions. Then students can receive appropriately scaffolded tasks and more targeted feedback, even if they are not aware of the learning progression but especially if they are (Rablin, 2024).

One potentially productive avenue of research and development has been made possible by advances in technology. Because of the internet, classroom walls are more porous than they once were, and this means that school-university partnerships that pair content-area teachers with external assessment researchers are possible in real time. Several programs of research have put together the formative benefits of a learning progression, the assessment design expertise of university researchers, and teacher management of classroom learning (e.g., Confrey & Toutkousian, 2019; Wilson & Draney, 2004; Wilson & Lehrer, 2021), as assessments designed outside the classroom are used inside the classroom, with information flowing both ways. A benefit of this kind of research is that it fosters advances in both learning progressions and assessment at the same time.

Section Summary

To say assessment serves learners when it is formative is almost a tautology, amounting to saying assessment serves learners when assessment informs their learning. Nevertheless, this seems to be a point that needs to be made (Shepard et al., 2018). This section drilled one step deeper, to show how assessment informs learners best when it helps them focus their attention on a learning goal and activate regulatory processes to move closer to it. Specific features of assessment that research has shown to facilitate this include designing high-quality assessment and learning tasks that match learning goals, activating the formative learning cycle in classroom instruction and assessment, providing students feedback that moves learning forward, and where possible using learning progressions to design and interpret instruction and assessment.

How does validation shift as the emphasis in assessment shifts to serving learners?

The phrase "to serve learners" is a purpose statement. Assessment purpose invokes validity because it has to do with inquiring into the appropriateness of interpretations and uses of assessment information (Kane, 2013, 2016). Making claims that students need assessment to function in certain ways and to have certain characteristics raises empirical questions that are at root validity questions.

Validation arguments (Kane, 2013, 2016) typically support interpretation and use (sometimes called meaning and impact, Lederman, 2023) in assessment. Several authors have pointed out that as the emphasis shifts from informing educators and administrators to serving learners, the emphasis in validation shifts proportionately

from interpretation/meaning toward use/impact (Hopster-den Otter et al., 2019; Kane & Wools, 2020; Lederman, 2023; Moss, 2016) and, indeed without assessment use score interpretation becomes a moot point (Sireci, 2016).

Lederman (2023) argued more emphasis on use or impact could be incorporated into validation work to disrupt assessment-based racial injustice. Because score meaning may differ for different groups, emphasizing impact is the key to pursuing racial justice in assessment. Some of the same arguments work for supporting learners in assessment, as well. Emphasizing use of assessment information increases the importance of learners' achievement and motivational outcomes relative to score meaning—in other words, it elevates the purpose of serving learners.

When validation expands into questions of use and impact, the contexts of the organizations that use data and the resources of their educational professionals become relevant sources of evidence for validation (Moss, 2016). Moss's argument about educators' use of tests in their schools and districts could easily be extended to students' use of assessments in their classroom learning cultures. If the claim is made that an assessment serves learners, the validity of the assessment information depends in part on students' knowledge and interpretation of that assessment information. For example, for a self-assessment used at a pause point in a classroom unit of instruction, students' understanding of the learning goal the assessment is meant to inform and the criteria by which they will know where they are and where to go next will affect the degree to which they can take the assessment information on board as feedback and use it productively. If students do not have a clear enough concept of what they are trying to learn, information about where they are now will have limited usefulness to them and thus limited validity for supporting them as learners.

Validation can regard students as learners, examinees, or contestants (Dorans, 2012). As assessment shifts from the decontextualized measurement characteristic of conventional large-scale tests to measurement contextualized as part of a student's formative learning cycle—that is, as assessments develop to better serve learners—students shift from being examinees to being learners and sometimes contestants. Dorans (2012) posits that in the case of large-scale assessment, from a contestant perspective—in addition of course to simply wanting to win—students would see assessment interpretation and use as valid if the assessment created

a fair race characterized by reliable outcomes, acceptable scoring, clear rules, and empirically verified interpretations. This chapter has shown that, in the case of classroom learning, from a learning perspective students would see assessment interpretation and use as valid if the assessment: (1) was situated in a supportive classroom learning culture that included honoring their own sociocultural context, (2) was situated in a lesson or series of lessons for which they understood clear learning goals and criteria for success, (3) was used as part of their participation in the formative learning cycle, and (4) provided feedback that moved learning forward and was coupled with an opportunity to do so. I would argue that for classroom formative assessment, where results do not need to generalize and often are not scores at all, the latter are the most important validation criteria.

The learner perspective on validity and the contestant perspective on validity are broadly comparable to mastery and performance goal orientations (Maehr & Zusho, 2009). Achievement goal theory posits two potential reasons why students may be motivated to learn. Briefly, students with a mastery goal orientation are motivated because they want to master the content (the learner perspective); students with a performance goal orientation are motivated to demonstrate to others that they are smart, or smarter than their classmates (performance-approach goals, the contestant perspective) or to avoid seeming not to be smart (performance-avoidance). Both mastery goals and performance-approach goals have small positive effects on academic achievement (Maehr & Zusho, 2009; Senko, 2019). Depending on the subject matter and specific classroom culture, students adopt either or both of these goal orientations. Therefore, it seems prudent that as the validation literature expands to include students' perspectives, both the learner and contestant perspectives should be considered.

Summary, Recommendations, and Conclusions

Summary and Principles

In this chapter, I have shown that first, learners need a supportive classroom learning culture that honors their own culture and leverages their own funds of background knowledge and experience. Second, learners need a clear understanding of learning goals and success criteria. Clear and explicit communication of learning goals can be a means for promoting equity because all students have access to the goal and criteria for good work, not just those whose background allows them to infer these things from a lesson where they are only

implicit. 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," is clearly implicated.

Third, learners need assessment to be situated in the formative learning cycle, in which students set their sights on a learning goal and actively pursue it using their cognitive, affective, and behavioral resources. Principle 3: "Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition," is clearly implicated. Fourth, as they are navigating the formative learning cycle, learners need effective, high-quality feedback and opportunities to use it. Principle 5, "Feedback, adaptation, and other relevant instruction should be linked to assessment experiences," is clearly implicated.

I have argued that to meet these learner needs, those who design and use assessments should give attention to the quality of assessment questions and tasks, the use of formative assessment processes and tools, the quality and usefulness of both scores and descriptive feedback that result from assessment, and the use of learning progressions where appropriate. The next section interprets these points into more focused recommendations for those who develop and use assessment to serve learners.

Recommendations for Developing Educational Assessments to Serve Learners

In deciding how to translate the chapter's discussion into recommendations, I had to make a grain size decision. In keeping with the size and scope of the chapter, I offer these recommendations at a middle grain-size level. For example, Recommendation 1 could say "More student voice," which might be too general to be helpful, or it could include a list of many different practical ways to incorporate student voice into the design, interpretation, and use of various different kinds of assessments in various subject matters and grade levels; a how-to list of this sort is definitely needed but is beyond the scope of this chapter. Thus, I offer these recommendations that, to me, follow from the chapter's discussion and are specific enough to at least move the conversation forward and suggest both future research and future professional development for those involved in assessment. Table 1 presents a list of assessment developments that, in my view, will enable the development of educational assessments that serve learners better in the future.

Table 1.
Developing Educational Assessments to Serve Learners

Recommendation	References
Better understanding of the results of increasing student voice and choice in assessment on the interpretation and use/impact of assessment results	Andrade et al. (2019); Heritage & Wylie (2020); Hood (1998); Ladson-Billings (2014); Randall (2021); Solano-Flores & Nelson-Barber (2001); Taylor & Nolen (2022)
Increase in the quality of assessment questions and tasks, both classroom and external	Brookhart & DePascale (in press); Carless (2015); Chen et al. (2017); Heritage & Wiley (2020); Wiggins (1998); Wilson (2004)
Increase in the quality and usefulness (to learners) of various kinds of scoring schemes and feedback comments and understanding of the conditions under which to deploy them	Hattie & Timperley (2007); Jonsson & Panadero (2018); Shute (2008); Van der Kleij, Feskens, & Eggen (2015)
Increase in the repertoire of available formative assessment strategies for both teachers and students	Bailey & Heritage (2018); Heritage & Wylie (2020); Brookhart (2017)
Increase in the use of classroom assessment results, especially classroom formative assessment, and a concomitant respect and understanding of the place of high-quality classroom assessment in the learning experience for learners	Safir & Dugan (2021)
More judicious use of external assessment results: using accountability and other summative assessment results only to raise questions about learning, not to answer them; developing more diagnostic external "formative" assessments, perhaps using emerging technology and measurement methods	Bradshaw & Levy (2019); Shepard et al. (2018); Zenisky et al. (in press)

Recommendation	References
More research and development on promising programs that mix internal, situated classroom work with external assessment, if and only if accompanied by the development of learning progressions and other tools, deeply criterion-referenced, and tied to students' learning experiences	Confrey & Toutkousian (2019); Wilson & Draney (2004); Wilson & Lehrer (2021)
Educative. Assessments build educator and student understanding of and experience with high-quality teaching and learning in the discipline.	The assessment tasks, student data, and supports for interpretation—should build educator understanding of what high-quality disciplinary teaching and learning look like, what kinds of tasks can develop and evaluate that learning, and how to provide feedback in ways that support progress toward these goals.
Assessments should attend carefully to the learning of teachers and students alike and are designed such that teachers also feel they learned something meaningful about their practice. What assessments signal, measure, and provide information about should directly speak to the actions and decisions we want students, educators, and leaders to make—and help them learn how to do so and why it is important. This may be accomplished by incorporating performance tasks into the instructional process; releasing items, tasks, and student work so that educators can see the kinds of tasks students are being asked to accomplish and what scores reflect; involving educators in designing and scoring tasks; providing task and student response annotations; providing concrete next steps to take, aligned to features of high-quality teaching and learning in science and based on student performance profiles; and making student experience data available to educators and leaders to contextualize performance.	

Each recommendation in the list is accompanied by some of the citations that inspired and support it. As the references show, some work at least at the concept level has begun for each of these recommendations, and in many cases research and development has begun, as well.

Conclusion

In this chapter, I have tried to shine a light on aspects of educational assessment that might help move the needs of learners closer to the center of assessment. Perhaps it should be surprising that considering the needs of learners has not always been the first principle of all educational assessment. However, it clearly has not (Dorans, 2012; Shepard, 2000). This chapter stands on the shoulders of others who would move the needs of learners into a central place in assessment, and I have tried to cite a wide variety of work to demonstrate that. Often that meant that large bodies of work were just mentioned and cited, or ideas that could be whole chapters in themselves just received a paragraph. I hope this chapter prompts readers to pursue these thoughts further.

References

Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition and Instruction, 37*(1), 1–37. https://doi.org/10.1080/07370008.2018.1539735

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261–271. https://doi.org/10.1037/0022-0663.84.3.261

Andrade, H. L., Bennett, R. E., & Cizek, G. J. (Eds.). (2019). *Handbook of formative assessment in the disciplines.* Routledge.

Andrade, H. L., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice, 27*(4), 350–372. https://doi.org/10.1080/0969594X.2019.1571992

Andrade, H. L., Brookhart, S. M., & Yu, E. C. (2021). Classroom assessment as co-regulated learning: A systematic review. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.751168

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation, 10*, 1–11. https://scholarsarchive.library.albany.edu/edpsych_fac_scholar/2

Bailey, A. L., & Durán, R. (2020). Language in practice: A mediator of valid interpretations of information generated by classroom assessments among linguistically and culturally diverse students. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement*. Routledge. https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9780429507533-4/language-practice-alison-bailey-richard-dur%C3%A1n

Bailey, A. L., & Heritage, M. (2018). *Self-regulation in learning: The role of language and formative assessment.* Harvard Education Press.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice, 38*(2), 79–88. https://doi.org/10.1111/emip.12247

Brookhart, S. M. (2017). Formative assessment in teacher education. In D. J. Clandinin & J. Husu (Eds.), *International handbook of research on teacher education* (pp. 927–943). Sage.

Brookhart, S. M. (2018). Summative and formative feedback. In A. A. Lipnevich & J. K. Smith, (Eds.), *The Cambridge handbook of instructional feedback* (pp. 52–78). Cambridge University Press.

Brookhart, S. M. (2024). Using rubrics in basic education: A review and recommendations. *Estudos em Avaliação Educacional, 35*, Article e10803. https://doi.org/10.18222/eae.v35.10803

Brookhart, S. M., & DePascale, C. A. (in press). Assessment to inform teaching and learning. In Cook, L., & Pitoniak, M. (Eds.), *Educational measurement* (5th ed.). Oxford University Press.

Brookhart, S. M., & McTighe, J. (2017). *The formative assessment learning cycle*. ASCD.

Chen, F., Lui, A. M., Andrade, H., Valle, C., & Mir, H. (2017). Criteria-referenced formative assessment in the arts. *Educational Assessment, Evaluation and Accountability, 27*, 297–314. https://doi.org/10.1007/s11092-017-9259-z

Confrey, J. (2019, May). *A synthesis of research on learning trajectories/progressions in mathematics*. OECD EDU/EDPC(2018)44/ANN3. https://www.oecd.org/education/2030-project/about/documents/A_Synthesis_of_Research_on_Learning_Trajectories_Progressions_in_Mathematics.pdf

Confrey, J., & Toutkoushian, E. (2019). A validation approach to middle-grades learning trajectories within a digital learning system applied to the "Measurement of Characteristics of Circles." In J. Bostic, E. Krupa, and J. Shih (Eds.): *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 67–92). Routledge.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4). 438–481. https://doi.org/10.3102/00346543058004438

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37. https://doi.org/10.1111/j.1745-3992.2012.00250.x

Esteban-Guitart, M., & Moll, L. C. (2014). Lived experience, funds of identity and education. *Culture & Psychology, 20*(1), 70–81. https://doi.org/10.1177/1354067X13515940

Garcia-Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology. 9*(3), 1043–1062.

Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165–189). Taylor and Francis.

Hattie, J. A. C. (2009). *Visible learning*. Routledge.

Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback*. Routledge.

Hattie, J., Crivelli, J., Van Gompel, K., West-Smith, P., & Wike, K. (2021). Feedback that leads to improvement in student essays: Testing the hypothesis that "Where to next" feedback is most powerful. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.645758

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heritage, M. (2013). Gathering evidence of student understanding. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 179–195).

Heritage, M., & Wylie, E. C. (2020). *Formative assessment in the disciplines*. Harvard Education Press.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *Journal of Negro Education, 67*(3), 187–196. https://doi.org/10.2307/2668188

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement, 56*(4), 715–732. https://doi.org/10.1111/jedm.12234

Ibarra-Sáiz, M. S., Rodríguez-Gómez, G., & Boud, D. (2021). The quality of assessment tasks as a determinant of learning. *Assessment & Evaluation in Higher Education, 46*(6), 943–955. https://doi.org/10.1080/02602938.2020.1828268

James, M., Black, P., Carmichael, P., Conner, C., Dudley, P., Fox, A., Frost, D., Honour, L., MacBeath, J., McCormick, R., Marshall, B., Pedder, D., Procter, R., Swaffield, S., & Wiliam, D. (2006). *Learning how to learn: Tools for schools.* 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

Jonsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 531–553). Cambridge University Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211. https://dx.doi.org/10.1080/0969594X.2015.1060192

Kane, M. T., & Wools, S. (2020). Perspectives on the validity of classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 11–26). Routledge. https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9780429507533-2/perspectives-validity-classroom-assessments-michael-kane-saskia-wools

Laal, M., & Ghodsi, S. M. (2012). Benefits of collaborative learning. *Procedia—Social and Behavioral Sciences, 31*, 486–490. https://doi.org/10.1016/j.sbspro.2011.12.091

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491. https://doi.org/10.3102/00028312032003465

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: A.k.a the remix. *Harvard Educational Review, 84*(1), 74–84. https://doi.org/10.17763/haer.84.1.p2rj131485484751

Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education, 36*(3), 242–254. https://doi.org/10.1080/08957347.2023.2214654

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K–12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Leighton, J. P., Chu, M-W., & Seitz, P. (2013). Errors in student learning and assessment. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 185–208). Information Age Publishing.

Lipnevich, A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*(720195). https://doi.org/10.3389/feduc.2021.720195

Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. CLASS Research Report No. 9. Center for Instructional Innovation, University of Nebraska-Lincoln. https://www.researchgate.net/publication/247291218_Providing_Feedback_in_Computer-based_Instruction_What_the_Research_Tells_Us

Maehr, M. L., & Zusho, A. (2009). Achievement goal theory: Past, present, and future. In K. Wentzel & D. Miele (Vol. Eds.), *Handbook of Motivation at School: Vol 1*, (pp. 77–104). Taylor & Francis.

McMillan, J. H. (2018). *Using students' assessment mistakes and learning deficits to enhance motivation and learning.* Routledge.

Mosher, F. A. (2022). *Learning progressions. Routledge Resources Online—Education.* https://doi.org/10.4324/9781138609877-REE115-1

Moss, C. M. (2022). Learning targets and success criteria. Routledge *Resources Online—Education.* https://doi.org/10.4324/9781138609877-REE39-1

Moss, P. A. (2016). Shifting the focus of validity for test use. Assessment in Education: Principles, Policy & Practice, 23(2), 236–251. https://dx.doi.org/10.1080/0969594X.2015.1072085

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. *Assessment & Evaluation in Higher Education. 38*, 919–940. https://doi.org/10.1080/02602938.2012.758229

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*(2), 226–245. https://doi.org/10.1080/01619561003688688

Panadero, E., & Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practice connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology, 11*(2), 551–576. http://dx.doi.org/10.14204/ejrep.30.12200

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self- and co-regulated learning. *European Journal of Psychology of Education, 34*(3), 535–557. https://link.springer.com/article/10.1007/s10212-018-0407-8

Rablin, T. (2024). *Hacking student motivation: 5 assessment strategies that boost learning progression and build student confidence*. Times 10 Publications.

Randall, J. (2021). "Color-neutral is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90. https://doi.org/10.1111/emip.12429

Resnick, L. B. (1987). Learning in school and out. *Educational Researcher, 16*(9), 13–20+54. https://doi.org/10.3102/0013189X016009013

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255. https://doi.org/10.3758/BF03194060

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

Safir, S., & Dugan, J. (2021). *Street data: A next-generation model for equity, pedagogy, and school transformation.* Corwin.

Senko, C. (2019). When do mastery and performance goals facilitate academic achievement? *Contemporary Educational Psychology, 59*. https://doi.org/10.1016/j.cedpsych.2019.101795

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Shepard, L. A. (2018a). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165–174. https://doi.org/10.1080/08957347.2017.1408628

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). 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. https://doi.org/10.1111/emip.12189

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226–235. http://dx.doi.org/10.1080/0969594X.2015.1072084

Slade, S., & Griffith, D. (2013). A whole child approach to student success. *KEDI Journal of Educational Policy, Special Issue,* 21–35. https://www.kedi.re.kr/eng/kedi/bbs/B0000005/list.do?menuNo=200067

Smarter Balanced Assessment Consortium. (n.d.). *Understanding the Smarter Balanced reporting system for educators.* https://youtu.be/sYMY4CJUO6g

Smith, J. K., & Lipnevich, A. A. (2018). Instructional feedback: Analysis, synthesis, and extrapolation. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 591–603). Cambridge University Press.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573. https://doi.org/10.1002/tea.1018

Stobart, G. (2018). Becoming proficient. In A. A. Lipnevich & J. K. Smith, (Eds.), *The Cambridge handbook of instructional feedback* (pp. 29–51). Cambridge University Press.

Taylor, C. S., & Nolen, S. B. (2022). *Culturally and socially responsible assessment: Theory, research, and practice.* Teachers College Press.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* Jossey-Bass.

Wiliam, D., Fisher, D., & Frey, N. (2024). *Student assessment: Better evidence, better decisions, better learning.* Corwin.

Wilson, M. (2004a). A perspective on current trends in assessment and accountability: Degrees of coherence. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 272–283). 103rd Yearbook of the National Society for the Study of Education, Part II. University of Chicago Press.

Wilson, M., & Draney, K. (2004b). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 132–154). 103rd Yearbook of the National Society for the Study of Education, Part II. University of Chicago Press.

Wilson, M., & Lehrer, R. (2021). Improving learning: Using a learning progression to coordinate instruction and assessment. *Frontiers in Education, 6*:654212. https://doi.org/10.3389/feduc.2021.654212

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Zenisky, A., O'Donnell, F., & Hambleton, R. K. (in press). Reporting scores and other results. In Cook, L., & Pitoniak, M. (Eds.), *Educational measurement* (5th ed.). Oxford University Press.

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and overview. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 1–12). Routledge.

Toward a Culturally Self-Regulated Dynamic Pedagogy Assessment System

Héfer Bembenutty

Abstract

This chapter introduces the Culturally Self-Regulated Dynamic Pedagogy Assessment System (CSDPAS), a comprehensive framework designed to integrate self-regulated learning (SRL) with culturally responsive pedagogy. Building upon Zimmerman's cyclical model of SRL and the Dynamic Pedagogy framework, the system emphasizes the interconnected roles of curriculum, instruction, and assessment to enhance academic outcomes for diverse learners. CSDPAS is grounded in the belief that embedding SRL strategies—goal setting, self-monitoring, and self-reflection—into culturally inclusive educational practices fosters student agency, equity, and academic success. The chapter outlines how teachers and students can engage in self-regulation across the phases of learning and teaching, supported by evidence-based practices that highlight the importance of self-efficacy, motivation, and culturally relevant assessments. Case studies demonstrate the model's effectiveness in improving student engagement, teacher satisfaction, and learning outcomes, particularly in diverse classroom environments. The author advocates for future research to expand this model and emphasizes the need for dynamic, culturally attuned assessment systems that promote lifelong learning and equitable education.

Self-regulated learning is a process that can benefit both teachers and students alike. *Self-regulated learning* refers to the ability of individuals to control their learning process by managing their thoughts, emotions, behaviors, and actions to pursue valuable academic outcomes successfully (Zimmerman, 2013). By purposefully pursuing academic goals, students acquire skills, improve their academic outcomes, and achieve better results. Conversely, teachers can use self-regulation in all aspects of their teaching profession (Bembenutty et al., 2015; DiBenedetto, 2018). It can also help parents and educators better understand and support their children's or students' learning progress. Self-regulated learning is a reliable approach that can enhance teaching, learning, and academic outcomes for everyone involved (Greene et al., 2024).

While there is extensive evidence that self-regulated learning is associated with valuable teaching, learning, and academic outcomes, much still needs to be discovered about how assessment can facilitate self-regulated learning. There is a need for a better understanding of how self-regulated learning can promote assessment equity, accountability, and adaptation in diverse classrooms with students and teachers from diverse backgrounds. It is also vital to understand how self-regulated learning can facilitate the transfer of knowledge and skills and the development of future goals and objectives. Despite its effectiveness, there is a lack of a comprehensive and dynamic pedagogical model that integrates curriculum, instruction, and assessment within the self-regulated learning framework (Bondie & Zusho, 2018; White & DiBenedetto, 2015). Developing such a model can produce positive academic outcomes for students' self-regulated learning and enhance teachers' ability to adopt effective curriculum, instruction, and assessment. This model can be particularly beneficial for learners and educators from diverse backgrounds who aspire to learn, teach, and assess in inclusive and equitable classroom environments, fostering a sense of belongingness and inclusivity (Armour-Thomas & Gordon, 2013, 2025).

In this chapter, I have five primary objectives. First, I present an overview of the self-regulated learning theory and five major hallmarks for learning-centered curriculum, instruction, and assessment. Second, I discuss how self-regulated learning is a theoretical foundation and guiding framework for understanding curriculum, instruction, and assessment processes. Third, I demonstrate how self-regulated learning is aligned with and supports the Dynamic Pedagogy framework (Armour-Thomas, 2017; Armour-Thomas & Gordon, 2013), which

integrates assessment with curriculum and instruction centered on learning while embedding equity, assessment, and cultural practice; and introduce the Culturally Self-Regulated Dynamic Pedagogy Assessment System. I describe how the three phases of self-regulated learning are theoretically construed and embedded in the transformational outcomes of rigor, love, freedom, and joy at each stage of the curriculum, instruction, and assessment processes. Fourth, I review evidence-based research to underscore the importance of self-regulated learning in promoting a diverse, equitable, and inclusive educational assessment system, drawing from the experiences of both preservice and in-service teachers. Finally, I recommend how self-regulated learning can enable equitable curriculum, instruction, and assessment practices in the 7–12 classroom. I conclude the chapter by emphasizing the imperative need for further research and practices promoting a culturally self-regulated educational assessment system, where feedback to learners and educators is essential to the formative assessment process. This comprehensive approach aims to enhance understanding and implementation of self-regulated learning within educational settings, contributing to developing a more inclusive and effective learning environment for all students.

Self-Regulated Learning

Self-regulated learning processes predict effective teaching and learning (Zimmerman, 2013). Self-regulated learning can transform how we approach curriculum, instruction, and assessment and predict effective teaching and learning (Bembenutty et al., 2013; Butler et al., 2017; Kitsantas et al., 2024). Self-regulated learning involves the acquisition of learning habits, study skills, learning strategies, and metacognitive skills associated with positive academic outcomes. Skilled self-regulated learners set academic goals, assess their motivation and task value, and evaluate and monitor their performance and outcomes. Self-regulated learning can help learners develop the necessary skills and strategies to achieve their academic goals more effectively.

Learners' Self-Regulated Learning

The first phase of self-regulated learning is forethought. During the forethought phase, learners set goals, plan strategies, and activate prior knowledge. They also assess their motivational beliefs, such as self-efficacy, outcome expectation, task value, and interest. For instance, goal setting involves identifying target outcomes linked to standards for assessing performance (White & DiBenedetto, 2018). Setting

specific, measurable goals focusing on short- and long-term outcomes is vital for successful self-regulated learning. It is also essential to have an acceptable level of self-efficacy beliefs to achieve these goals. Self-efficacy is a motivational component of self-regulated learning that positively predicts performance.

Self-efficacy refers to individual beliefs about their capability to perform designated tasks (Bandura, 1997). Self-efficacy effectively predicts students' motivation and learning, interacts with self-regulated learning processes, determines activity choices, effort, persistence, and emotional reactions, and mediates academic achievement (Zimmerman, 2000). Like self-regulated learning, self-efficacy is assessed through performance capabilities rather than personal qualities, such as physical or psychological characteristics (Zimmerman, 2000). Understanding the importance of self-regulated learning and self-efficacy can help learners achieve their academic goals more effectively.

In the performance phase, learners monitor their progress, apply strategies, and seek help. This phase is where self-control becomes crucial, and learners must activate attention focusing, self-administration of instruction, and enacting task analysis. Self-observation is also a key aspect of this phase, and learners can benefit from self-recording and self-monitoring tasks or thoughts. Equally important is the role of help-seeking from knowledgeable sources, such as teachers, advanced peers, or multimedia outlets. When learners encounter academic challenges they believe are difficult, seeking help from the teacher as a self-regulated strategy can be highly beneficial. This instrumental help-seeking approach is a key factor in promoting effective learning. In contrast, executive help-seeking, which involves learning avoidance or asking for solutions to tasks without fully understanding them, is less effective in promoting learning (Karabenick & Gonida, 2018).

In the self-reflection phase, learners assess their outcomes, recognize their strengths and challenges, and adapt their goals and strategies for future learning. This phase is where self-judgment of tasks occurs, involving self-evaluation and causal attribution. It also includes an assessment of self-satisfaction and adaptation to new situations. Reflecting on the cause of outcomes is crucial in this phase, as students who attribute positive results to appropriate strategic usage tend to remain focused on identifying strategies that will produce valuable outcomes. On the other hand, students who attribute positive outcomes to luck

may not put in the necessary effort in the future. Self-regulated learning can benefit learners as it enables them to plan, monitor, and evaluate their learning processes, goals, and strategies (Schunk & Greene, 2018; Zimmerman, 2013).

As Figure 1 displays, Zimmerman's cyclical model of self-regulated learning consists of three phases that influence each other: forethought, performance, and self-reflection.

- *Forethought Phase*: In this phase, learners establish specific, realistic, feasible, challenging, and attainable goals and strategies. They also identify their outcome expectancy, self-efficacy, and interest levels in reaching those goals and strategies.
- *Performance Phase*: This phase is crucial in the development of self-control and self-observation. Learners create positive images and outcomes of the task, stay task-focused, provide self-instruction, and monitor tasks strategically. Additionally, they engage in self-recording and self-experimentation.
- *Self-Reflection Phase*: After completing the task, learners enter the self-reflection phase. Here, they assess the results of their actions, gauge their satisfaction, identify the causes behind the outcomes, and modify their goals and strategies as needed. This phase acts as a feedback loop, enabling learners to improve and prepare themselves for future cycles of tasks.

Self-regulated learning has immense potential to provide a new perspective and vision for curriculum, instruction, and assessment as a dynamic pedagogical model that can address the challenges and opportunities of teachers and learners (Schunk & Greene, 2018). It is a valuable tool that can help teachers design, implement, and adjust their curriculum, instruction, and assessment practices to meet their students' diverse needs and preferences. Self-regulated learning is essential for successful assessment, as stated by Artzt and her associates, "Taking personal responsibility and control of one's learning is a hallmark of academic excellence. A critical factor in this type of learning that researchers define as self-regulated… is self-assessment" (Artzt et al., 2015, p. 8). As agentic individuals, teachers can be proactive and self-directed while pursuing valuable academic goals and engage in self-regulation and coregulation (Greene et al., 2024). The cyclical phases of self-regulated learning also apply to teachers, which means teachers can benefit from the same self-regulatory processes they instruct their students (Pape et al., 2013;

White, 2017; White & DiBenedetto, 2015). Kramarski and Kohen (2017) highlighted the dual self-regulation roles of teachers, emphasizing the need for suitable assessment methods to capture the dynamic, complex, and cyclical nature of self-regulation within the teaching and learning process.

Teachers' Self-Regulated Learning

Self-regulated learning can help teachers better understand their strengths and areas for improvement and adjust their teaching practices. It can be a powerful tool in enhancing teachers' and students' teaching and learning experiences. The forethought phase is crucial for teachers as they are proactive agents who generate goals, engage in strategic planning, activate intrinsic motivation and maintain self-efficacy for learning and teaching. Teacher self-efficacy significantly shapes their thoughts, actions, lesson plan preparation, curriculum development, instruction, and assessment (Hoy et al., 2009). Teachers with high self-efficacy beliefs are also more effective in class management, teaching strategies, rapport with students, and effective assessment (Woolfolk et al., 2006). This is particularly important in challenging classroom situations, such as low student motivation, classroom management, unsupportive parents, and complex administration. Teacher self-efficacy empowers them to put effort and persistence into valuable teaching tasks, directing their actions and plans. This human agency is crucial in helping teachers navigate demanding situations and succeed in their profession (Bandura, 2006; Hoy et al., 2009).

There are four primary sources of self-efficacy: mastery experiences, vicarious experiences, verbal persuasion, and physiological and emotional states. Mastery experiences are the most powerful source of self-efficacy, as they involve direct personal success or failure in each domain. When teachers overcome challenges or achieve goals, they enhance their competence and confidence. Conversely, when people fail or encounter difficulties, they may lower their self-efficacy unless they attribute the failure to external or controllable factors. Vicarious experiences are the second source of self-efficacy, as they involve observing others perform a task or cope with a situation. When people see someone like themselves succeed or fail, they may infer that they can or cannot do the same. Verbal or social persuasion is the third source of self-efficacy, as it involves receiving encouragement or discouragement from others. When people are praised, supported, or motivated by someone they trust or respect, they may increase their self-efficacy.

Conversely, when people are criticized, doubted, or discouraged by others, they may decrease their self-efficacy. Verbal persuasion can help people overcome self-doubt and focus on their strengths and abilities. Physiological arousal or emotional states are the fourth source of self-efficacy, as they involve interpreting one's bodily and affective reactions to a task or situation. When people experience positive emotions, such as excitement, joy, or pride, they can significantly boost their self-efficacy. This understanding can instill a sense of optimism and confidence in teachers, knowing that their emotional state can significantly influence their self-efficacy (Bandura, 1997).

The literature supports the importance of teacher self-efficacy for a successful and healthy teaching career. Täschner et al. (2024) conducted a systematic review and meta-analysis of intervention studies promoting teacher self-efficacy. They analyzed over 115 studies, which included more than 11,284 pre-service and in-service teachers. The findings revealed interventions had a significant positive effect on promoting teachers' self-efficacy. Additionally, they found that interventions targeting mastery experiences were the most successful for pre-service teachers when examining the four sources of self-efficacy identified by Bandura (1997).

In the performance phase of learning, teachers can control their motivation and emotions, use effective learning strategies, seek help when required, and activate their metacognitive skills to ensure successful task completion and positive outcomes. While metacognition and self-regulation are used often interchangeably, they emphasize distinct aspects of learning. Metacognition involves thinking about cognition and cognitive structures, while self-regulated learners focus on regulating the behavior, cognition, feelings, and actions related to the learning process and outcomes. However, there is a debate about whether self-regulation is a subordinate component of metacognition. Regarding classroom assessment, Armour-Thomas adopted the notion that self-regulation is a subordinate component of metacognition. Regardless of this debate, it is crucial to understand that effective self-regulated teachers skillfully use metacognitive skills by planning, controlling, and monitoring their cognitive processes, leading to better learning outcomes. Therefore, it is essential to prioritize the development of self-regulated learning skills in teachers, as this will help them become more effective in their roles. By mastering the art of self-regulation, teachers can ensure positive classroom experiences for their students

and better learning outcomes. In the self-reflection phase, teachers assess their satisfaction with task completion and self-evaluate outcomes, examine their attributions and self-reaction to outcomes, and adapt their performance. This emphasis on self-reflection can make teachers feel more introspective and self-aware, enhancing their professional growth and effectiveness.

Integration of the Dynamic Pedagogy and Self-Regulated Learning

Self-regulated learning and the Dynamic Pedagogy framework aim to improve students' learning and teachers' ability to design and implement curriculum, instruction, and assessment. In this chapter, the focus is on breaking down silos (Matthews & Wigfield, 2024) by integrating the Dynamic Pedagogy and self-regulated learning frameworks into the *Culturally Self-Regulated Dynamic Pedagogy Assessment System*. This integrated approach emphasizes curriculum, instruction, and assessment while considering the cultural endeavors of both teachers and students. The literature supporting both models is vast and highlights the potential of each approach. Self-regulated learning is an essential component of the dynamic system, as Kaplan, Neuber, and Garner (2017) described. It encompasses content and strategic knowledge and considers the influence of culture, social context, subject domain, and the individual's implicit dispositions. Their dynamic pedagogy emphasizes the interconnectedness of several factors in shaping an individual's learning process and underscores the importance of self-regulation in achieving academic success.

The Dynamic Pedagogy framework has made significant strides in providing empirical evidence and conceptual integration (Armour-Thomas, 2008, 2017; Armour-Thomas & Gordon, 2013). However, self-regulated learning has also progressed in recent years, particularly emphasizing instruction, assessment, and students' learning (Cleary & Russo, 2024; Schunk & Greene, 2018). Although curriculum and assessment have only sometimes been at the forefront of the self-regulated learning approach, this model is consistent with and can support the Dynamic Pedagogy framework. Both models integrate assessment with curriculum and instruction centered on learning, emphasizing equity, assessment, and cultural practice. The self-regulated learning processes and the dynamic pedagogy framework are interconnected and can work concomitantly to enhance teachers' and students' teaching and learning experiences.

The Dynamic Pedagogy framework developed by Armour-Thomas and Gordon (Armour-Thomas & Gordon, 2013, 2025) is a powerful approach to teaching that emphasizes the integration of curriculum, instruction, and assessment to enhance learning outcomes (See Figure 2). In this approach, the key to dynamic pedagogy lies in the interconnection between these three elements, which includes adaptation and response to learners' behavior. In this context, pedagogy refers to the process and outcomes of student learning resulting from effective curriculum, instruction, and assessment. They distinguished pedagogy from instruction. Instruction refers to specific approaches teachers use to promote learning, while pedagogy is an umbrella term encompassing all three elements and how they work together to promote learning.

As a rationale for learning-centered assessment within the Dynamic Pedagogy framework, Armour-Thomas and Gordon argue that if the goal is to understand students' learning about determined standards, then assessment should not function separately from curriculum, as they both play a crucial role in understanding students' knowledge about determined standards and principles. They also posited that assessment could serve as a valuable feedback loop for instruction, allowing teachers to understand their strengths and areas for improvement, which could lead to more effective teaching practices and improved student learning outcomes.

The Venn diagram representation of the Dynamic Pedagogy model developed by Armour-Thomas and Gordon (2013) illustrates the interconnected relationships between curriculum, instruction, assessment, and learning, with the latter being the ultimate focus (See Figure 2). The nine dimensions of learning outcomes centered on the learners are fascinating, as they emphasize the importance of prior knowledge, social context, and metacognitive competence in the learning process. The model recognizes all children's potential to learn and the importance of meaningful learning that involves transferring knowledge to other contexts.

The Dynamic Pedagogy model is a valuable framework for teachers in designing effective curriculum, instruction, and assessment practices that promote student learning outcomes. The nine dimensions are consistent with the perception of learning within the self-regulated learning approach. Learning is construed as a function of the interrelation between the individual, the environment, and the behavior produced by the individual and the context (Bandura, 1997). However,

learning is not determined by external stimuli of reinforcement or punishment, nor by intrapsychic thoughts or experience. Learning is a function of the individuals' self-beliefs, agentic capabilities, forethought, and execution of actions. It also involves the capacity to plan, monitor, and control thoughts and actions, as well as self-reaction and self-reflection. From the self-regulated learning cyclical process, learning comprises the ability to set goals, plan, plan actions, monitor progress while reaching objectives, and reflect on outcomes.

The curriculum Dynamic Pedagogy strand covers the ideas, rules, criteria, and resources teachers use to facilitate learning. It also encompasses the content knowledge domain and how knowledge is arranged, built, and communicated to learners (See Table 1). Effective curriculum is delivered at a suitable level, with a logical sequence and appealing features that appropriately draw students' attention and relate to them. The link between curriculum and assessment is based on the idea that the choice of curriculum tools should align well with the evaluation of the student's learning outcomes, and assessment should be limited to only the content of the curriculum taught to the students. Self-regulated learning is embedded within the curriculum dimension of Dynamic Pedagogy's Venn diagram (See Figure 2). At the macro level, the curriculum is represented by Armour-Thomas and Gordon in a large shape. At the micro level, self-regulated learning is displayed by three small cycles within the large curriculum shape, representing the three cyclical phases: forethought, performance, and self-reflection. The curriculum design and implementation should be guided by the principles of cultural self-regulated pedagogy, which aims to foster self-regulated learning among diverse learners.

As discussed earlier, the culturally self-regulated pedagogy involves three cyclical phases: forethought, performance, and self-reflection. In the forethought phase, the curriculum should provide clear learning goals, expectations, strategies, and resources for planning and self-motivation. In the performance phase, the curriculum should offer a variety of media and formats to deliver the content and opportunities for students to seek feedback and monitor their progress. In the self-reflection phase, the curriculum should include tools and activities that help students evaluate their learning outcomes, reflect on their attributions and adaptability, and assess their self-satisfaction and self-efficacy.

During self-reflection, the curriculum should encourage students to set new goals, adjust their strategies, and celebrate their achievements. By following this

pedagogical approach, the curriculum can address the needs of all learners, especially those from minoritized and diverse backgrounds. The curriculum can also promote co-regulation between teachers and students as they share their thoughts, emotions, and actions related to the learning tasks (Greene et al., 2024; Hadwin et al., 2017). The cultural self-regulated pedagogy supports a proactive curriculum fostering self-fulfilling academic self-regulation cycles (White & Bembenutty, 2014), which can lead to effective and meaningful learning within a diverse curriculum (Artzt et al., 2015; White & Bembenutty, 2014).

The instructional Dynamic Pedagogy strand consists of strategies helpful to facilitate learning, including guided practice, supervised independent practice, modeling, scaffolding, and peer learning. This strand is related to assessment by revealing strengths and limitations in the assessment process. Given the assessment feedback, teachers can implement new instructional approaches that could result in effective learning. Self-regulated learning is embedded within the instruction dimension of Dynamic Pedagogy. Instruction is depicted in an oversized shape, and three small cycles within the large instruction shape display the self-regulated learning processes.

In the forethought phase, teachers can create opportunities for students to self-assess their self-efficacy beliefs, interest, and task value. Teachers can model ways to set measurable, realistic, and manageable goals and assist students in identifying their learning objectives and strategies. In the performance phase, teachers can help students self-monitor their progress by providing self-monitoring forms or logs and inviting them to seek help without concerns about being perceived as highly dependent. Teachers can invite students to assess their self-efficacy again to see whether it has fluctuated as they remain goal oriented.

In the self-reflection phase, teachers can ask students to engage in self-assessment or practice peer assessment and self-evaluation and help them adopt appropriate attributions for academic success or failure. It is essential to have a culturally self-regulated pedagogy in the classroom. Both teachers and students can be proactive, agentic, intentional, and self-directed, willing to engage in socially shared regulation and coregulation while embracing equity and diversity. Effective classroom instruction depends on orchestrating the needs of both students and teachers. The instruction is shaped by the teacher's agency, self-efficacy beliefs, and self-reflection on performance. Similarly, students' learning is influenced by

their agency, thoughts, self-efficacy beliefs, self-regulatory competencies, and the classroom context.

The assessment strand of Dynamic Pedagogy, with its two components: online probes and metacognitive probes, plays a pivotal role in promoting student learning and understanding. The *online probe* component helps teachers assess students' prior knowledge, skills, and readiness for new learning, aiding in identifying misconceptions and ensuring students have acquired the necessary knowledge and skills. In this context, the term "online" does not pertain to its conventional association with technology or digital platforms. Instead, it refers explicitly to real-time, interactive assessments of students' understanding during the learning process. These assessments, often conducted through questioning, involve students responding to open-ended tasks in a live, immediate manner. This approach aligns with the concept of "learning probes" as described by Slavin (2018), where educators gauge comprehension and engagement dynamically within the instructional environment.

While online probes can leverage technological tools such as computers and social media platforms (Golmohammadi, 2022), their core purpose remains rooted in fostering active participation and deeper cognitive engagement during the learning experience. The *metacognitive probe* component helps students become aware of effective learning strategies and how they can be applied to enhance their learning. Jenkins and Shoopman (2019) examined college students' misconceptions when molecular orbital diagrams are commonly taught and used for describing chemical bonding. Written probes were used to identify misconceptions, and it was found that many struggled to use and interpret the diagrams. They observed that metacognitive probes, like written probes, help calibrate students' comprehension. The assessment strand is interconnected with the curriculum strand, ensuring that the assessment is linked to the content covered in class.

Feedback plays a vital role in this strand, impacting the content and adaptation of the curriculum. Assessment is a critical component of fostering self-regulated learning and culturally self-regulated pedagogy. Self-regulated learning is ingrained within the assessment dimension of Dynamic Pedagogy, which operates at a macro level, as shown in the Venn diagram (See Figure 2), with a large shape. However, self-regulated learning operates at the micro level (represented by three small cycles within the large assessment shape) through three cyclical phases:

forethought, performance, and self-reflection. In the forethought phase, teachers ensure the assessment has undergone a rigorous task analysis, activated prior knowledge, and enabled students to use strategies within reasonable self-efficacy beliefs. In the performance phase, the assessment allows students to successfully apply strategies to complete the tasks. In the self-reflection phase, assessment serves as a tool for self-evaluation that provides feedback to learners about appropriate learning approaches and conveys expectations that learning is possible with acquired skills and effort. Regarding culturally self-regulated pedagogy, the assessment models of strategic learning offer opportunities for diverse ways of responding, are culturally fair, are sensitive to cultural diversity, and are administered fairly.

Within the assessment system, culturally self-regulated pedagogy (CSP) represents a comprehensive educational approach integrating self-regulated learning principles with cultural awareness, identity, and values. To illustrate, goal setting and self-efficacy are two culturalized processes and essential components of the CSP. To this point, Schunk and DiBenedetto have emphasized that "although goal setting may be universal, the types of goals set and how they are set are undoubtedly subject to cultural influences" (Bembenutty et al., 2023, p. 27). Similarly, they note, "Like goal setting, self-efficacy seems to represent a universal construct but is affected by cultural standards" (Bembenutty et al., 2023, p. 28). These observations highlight the need to integrate culturally self-regulated practices in diverse educational contexts to ensure that these processes align with students' cultural backgrounds. By doing so, educators can create more inclusive and effective learning environments that encompass all aspects of curriculum, instruction, and assessment.

Assessment is sensitive to bias and stereotypes. In his memoir, Edmund W. Gordon's (2014, Vol. I, p. 218) reflections underscore the impact of bias and stereotypes in assessments, particularly through the phenomenon of stereotype threat, as demonstrated by Steele and Aronson's study (Steele & Aronson, 2000). Their research revealed that minoritized college students' test performance could be influenced adversely by their awareness of societal perceptions labeling them as intellectually inferior. To Gordon, this critical finding highlights the need for equitable approaches in psycho-educational measurement. Gordon, drawing from such evidence, has been a strong advocate for more inclusive and fair assessment practices. His work has significantly informed and enriched the development of

the CSP, which aims to address systemic inequities in education within the self-regulated learning framework. The CSP emphasizes creating learning environments that respect and integrate diverse cultural experiences, fostering both equity and empowerment for all learners while focusing on promoting self-efficacy beliefs, enacting goals, agency, and self-reflection. Through his lifelong dedication, Gordon has contributed to advancing educational practices that prioritize fairness and cultural sensitivity, paving the way for more just and effective systems of evaluation and instruction. His efforts remain instrumental in shaping frameworks that challenge bias and promote inclusivity in education.

Unlike a mere adaptation of Culturally Responsive Teaching (CRT; Gay, 2018), or Culturally Relevant Pedagogy (Ladson-Billings, 2021), the CSP combines cognitive, metacognitive, and cultural strategies to create an inclusive learning environment supporting diverse students. CSP emphasizes empowering students to take ownership of their learning process by setting goals, monitoring progress, and refining strategies. It fosters essential skills such as time management, academic delay of gratification, critical thinking, and self-efficacy beliefs while embedding cultural relevance into the educational experience and providing a culturally valid and reliable curriculum and assessment. By incorporating students' cultural contexts and subjective experiences, CSP makes learning more meaningful and engaging. This framework values cultural diversity and equips learners with the ability to adapt their self-regulatory strategies to align with their unique cultural identities. The goal is to promote inclusivity and ensure that education is accessible and relevant for all students, enhancing their academic success, personal growth, and proactive self-regulation.

In contrast, Culturally Responsive Teaching (CRT) emphasizes integrating students' cultural identities into all aspects of education to enhance engagement and understanding. It seeks to make learning more relevant and effective for students from diverse backgrounds by valuing their cultural references. CRT employs teaching methods that respect and incorporate cultural diversity to boost student motivation and participation by making lessons relatable. This approach prioritizes equity and inclusion, addressing educational disparities by recognizing the significance of cultural diversity in the classroom. Teachers are encouraged to be aware of and responsive to students' cultural contexts, utilizing culturally relevant materials and examples within the curriculum. Collaboration with families and communities is also key to meeting cultural and academic needs. By fostering

an inclusive environment, CRT supports students in achieving academic success while affirming their cultural identities. This approach underscores the importance of creating a learning experience that values diversity and promotes meaningful connections between students' backgrounds and their educational journey.

Nevertheless, CSP and CRT both aim to create inclusive learning environments that honor students' cultural identities. However, their approaches differ in focus and implementation. CSP integrates SRL principles with cultural values, emphasizing the development of self-regulation skills in students while addressing their academic and cultural needs. Teachers in CSP act as facilitators, fostering proactive and agentic learning within a culturally relevant framework. In contrast, CRT emphasizes making education culturally relevant and equitable by incorporating cultural references into teaching strategies. While CRT focuses on creating a responsive environment, CSP goes further by proactively combining these principles with SRL to engage and motivate learners from diverse backgrounds actively. Both approaches aim to foster engagement, motivation, and academic success for culturally diverse learners. Educators can create a dynamic learning environment that respects cultural backgrounds while encouraging self-regulation and autonomy by integrating SRL with CRT principles. This dual approach ensures that students feel included and are empowered to take charge of their learning journey.

CSP and CRT both emphasize active student engagement. In CSP, students take ownership of their education by setting academic goals, monitoring progress, engaging in academic delay of gratification, assessing their level if self-efficacy beliefs, and adjusting strategies. They draw on their cultural knowledge to deepen understanding and adapt their learning approaches based on personal and cultural contexts, fostering self-motivation and agentic accountability. In contrast, CRT encourages students to actively contribute by sharing their cultural experiences, reflected in the curriculum and teaching methods. This approach enhances student engagement and motivation while promoting collaboration among peers and teachers to explore diverse cultural perspectives. CRT creates an inclusive learning environment that values and acknowledges students' cultural identities. Both frameworks aim to empower students by recognizing and leveraging their cultural backgrounds, fostering a sense of belonging, and enhancing learning outcomes through meaningful engagement.

Table 2 highlights the distinctions in *curriculum approaches* and roles between CSP and CRT. For CSP, the teacher's primary objective is to promote self-regulation skills and cultural awareness, while CRT emphasizes fostering cultural awareness and respect for diverse cultural backgrounds. From the students' perspective, CSP encourages goalsetting and planning with a focus on self-regulation, whereas CRT aims to ensure students see their cultural identities represented in the curriculum, fostering a sense of belonging and relevance. Table 3 provides a comparison of instructional approaches between CSP and CRT, illustrating how each framework approaches instruction differently, tailoring both teaching strategies and student engagement to align with their respective goals.

Table 4 displays differences in *assessment approaches* between CSP and CRT. In CSP, teachers emphasize formative feedback aimed at fostering students' self-regulation skills and encouraging them to refine their learning strategies to help students build content knowledge while promoting independent learning practices. In contrast, CRT focuses on providing culturally sensitive feedback that validates and acknowledges students' cultural identities, which is designed to support students' academic growth while affirming their cultural backgrounds, creating a more inclusive and supportive learning environment. For students, CSP assessments are centered on developing self-regulation and content mastery through iterative feedback. Meanwhile, CRT assessments prioritize recognizing and incorporating cultural identities into the learning process, ensuring that feedback aligns with students' cultural contexts to enhance their academic success. Both approaches aim to support student development, albeit through distinct lenses.

Research Evidence

Several studies and theoretical frames support integrating self-regulated learning within a dynamic pedagogy framework. Studies have shown self-regulated learning is associated with curriculum, instruction, and assessment. For instance, Bembenutty and Hayes (2018) conducted a study in an alternative learning center, which caters to middle and high school students assigned there for several reasons, such as suspensions or severe misconduct behaviors. These behaviors include drug use, fighting, sexual abuse, and delinquency, leading to a diverse student population with varying academic abilities. Some students were found to be reading at the third-grade level, highlighting the challenges faced by the educators in addressing the educational needs of such a heterogeneous group. The project's primary objective was to implement the culturally self-regulated dynamic

pedagogy assessment model aimed at introducing students to self-regulated learning through learners' self-assessment during instruction. This approach sought to empower students to take ownership of their learning process, thereby promoting a sense of accountability and autonomy.

Drawing from Zimmerman's self-regulated model, students actively engaged in a three-phase self-monitoring process during the lesson. In the forethought phase, which spanned the initial five minutes of the lesson, students delineated their objectives and outlined strategies for achievement. They gauged their self-efficacy and interest in the upcoming material. Throughout the lesson, in the performance phase, students continuously monitored their progress, evaluated their willingness to delay gratification by deferring immediate rewards, and assessed their self-efficacy levels. The culmination of the lesson involved the self-reflection phase, during which students appraised their satisfaction with their performance, made attributions for their outcomes, and devised plans for subsequent tasks or adjustments for unexpected outcomes. Concurrently, the teacher actively participated in these phase processes, serving as a model and providing scaffolding for students to co-regulate their performance. The teacher's ability to modify instruction based on student performance underscores the adaptive nature of this approach. Following in-class instruction, students were tasked with utilizing a homework log to self-monitor their completion of assignments. The homework log mirrored the three cyclical phases employed during in-class activities. Subsequently, students submitted their completed homework alongside the corresponding logs during the subsequent class session.

The results of Bembenutty and Hayes' (2018) study indicate the students demonstrated a prominent level of motivation and engagement with the self-monitoring form and homework log. Motivation and engagement were reflected in their interest, self-efficacy, willingness to delay gratification, ability to engage in self-assessment, and the teacher's positive performance assessment. The teacher reported a keen sense of satisfaction and motivation with the outcomes, highlighting the positive impact of integrating curriculum, instruction, and assessment on student academic achievement and teacher satisfaction. By incorporating self-regulated learning strategies into the instructional framework, the researchers aimed to foster a more inclusive and supportive learning environment conducive to the diverse needs of the student body. Thus, a significant outcome of this study was the ability of the self-monitoring form and the homework log to allow

students to express their goals and strategies based on their cultural background, self-identity, experience, and interests. This outcome underscores the importance of recognizing that curriculum, instruction, assessment, and self-regulated learning are all cultural enterprises that can favorably impact the teaching and learning processes, and incorporating students' cultures can positively impact the teaching and learning processes. These results emphasize the interconnected nature of curriculum, instruction, and assessment and their potential to support academic achievement and create a more culturally inclusive learning environment. It is evident that when these elements are effectively integrated, they can contribute to student success and teacher fulfillment. Students were able to return to their regular classrooms.

Bembenutty, White, and Velez (2015) illustrated how self-regulated learning produces positive educational outcomes when ingrained into curriculum, instruction, and assessment. Study participants were teacher candidates from minoritized backgrounds whose learning and teaching experience was transformed when their teacher educators introduced them to self-regulated learning. The teacher candidates experienced personal and academic challenges and, at some points, were at risk of academic failure. They did not know how to set goals, assess their self-efficacy beliefs, or identify effective learning strategies. Their help-seeking approaches were primarily avoidance or dependency and were ineffective in monitoring their learning and self-reflection. However, the teacher educator successfully integrated self-regulated learning into their curriculum, instruction, and assessments, positively impacting the teacher candidates. The teacher educators revised their traditional curriculum by ingraining into its self-regulated learning components, including self-efficacy and delay of gratification. For instance, the curriculum design added reading materials related to self-regulation. It required that the instruction and assessment involved be presented with language and rubrics reflecting strategic learning. The instruction was transformed in ways that reflected more like an academic. The educators modeled goal setting, motivation, and self-reflection during each instructional time and student teaching. The assessment process involved the triangulation of data sources, which included observation, questionnaires, self-reflections, and interviews for two years while considering the students' cultural background.

Bembenutty, White, and Velez's (2015) revealed a significant improvement in the students' self-regulation, as evidenced by various indicators such as heightened

teacher self-efficacy, a greater willingness to delay gratification, increased intrinsic motivation, and an increased sense of perceived responsibility. Through interviews, students expressed their enhanced preparedness for teaching and their positive outlook on their future careers in education. They also reported increased self-efficacy for learning and deliberate use of self-regulated learning strategies, further supported by faculty observations during their student teaching experiences. For instance, one of the students articulated,

> *I engage in time management. I have to make decisions about spending time with friends or getting my lesson plans done. My attitude in the classroom is positive. I push myself to be positive so the students can have a positive learning environment… I establish new goals for myself and my students. By sharing my goals with them it helps them to grow. I use post-test assessments to re-evaluate my whole lesson.* (Bembenutty et al., 2015, p. 65)

Bembenutty, White, and Velez's (2015) findings highlight the significant strides made by the students in terms of their self-regulation and preparedness for the teaching profession. They demonstrated a proactive approach to effectively managing their responsibilities, cultivating a positive learning environment, and establishing meaningful objectives for their development and that of their students. These findings not only signify the students' personal growth, but also underscore the potential impact of their future contributions to the field of education. The student's commitment to their growth and the cultivation of a supportive learning environment bodes well for their future success as educators, and their dedication serves as a testament to their readiness to influence the lives of their future students positively. By providing students with opportunities to set goals, assess their motivation, monitor their performance, and reflect on outcomes, they became more self-directed learners who could better manage their learning. Regular assessment and feedback also helped students identify their strengths and weaknesses and adjust their learning strategies.

In a recent study, Bembenutty (2023) assessed how integrating self-regulated learning and digital technologies can improve teaching practices in diverse postsecondary learning contexts. Teacher candidates were trained to recognize the value of self-regulated learning and technology for enhancing their proactivity, self-direction, and self-efficacy. The study aimed to foster teacher candidates' agency in pursuing their teaching career during their training programs and to promote a

culturally self-regulated pedagogy. In their educational psychology course, teacher candidates learned about self-regulatory processes and integrating digital technology into the curriculum. They learned how to become self-regulated learners and self-efficacious practitioners as they acquired knowledge and skills for teaching and fostering self-regulation among their future students. Teacher candidates developed a technology presentation in which they chose a technological tool to support instruction and learning. They explained how it could enhance self-regulation and address diverse learners' needs. They used various computer programs. One student who used Quizizz (https://quizizz.com/) for instructional purposes noted that it could help create class assignments, quizzes, pre-test reviews, and formative assessments (Bembenutty, 2023). Another who used Socrative (https://www.socrative.com/) observed that it could help assess prior knowledge, generate questions, monitor comprehension, and boost self-efficacy (Bembenutty, 2023). These examples show curriculum, instruction, assessment, and self-regulation integration.

Chen and Bonner (Bonner & Chen, 2019; Chen, 2023; Chen & Bonner 2020, 2023) developed a comprehensive framework integrating classroom assessment practices and self-regulated learning theory to facilitate academic growth and instruction. Following Zimmerman (2013), this framework consists of three main phases—forethought, performance, and self-reflection—and encompasses four stages of classroom assessment: pre-assessment, the cycle of learning, doing and assessing, formal assessment, and summarizing assessment evidence. The model emphasizes the activation of self-regulated learning at each stage, highlighting the dynamic interaction between assessment and self-regulated learning for both teachers and students, leading to effective classroom assessment.

During the forethought phase, students are encouraged to consider their prior experiences and individual differences while teachers gather information on students' prior attributes. This phase sets the stage for understanding the diverse needs of students and tailoring instruction accordingly. In the performance phase, students self-check while teachers monitor instructional checkpoints, creating an informal performance interactive assessment. Subsequently, during formal assessment, students continue to perform and self-check while teachers interpret and infer the results. This stage formally evaluates students' progress and understanding, informing future instructional decisions. Finally, in the summary of evidence and formal self-reflection phase, students are prompted to self-reflect and make

attributions while teachers make judgments and record outcomes. This phase encourages students to take ownership of their learning and allows teachers to assess the overall effectiveness of their instructional strategies. By incorporating self-regulated learning at each assessment stage, teachers can support students in developing essential skills such as goal setting, self-monitoring, and reflection.

Chen's (2023) study on the interactions between self-regulated learning and assessment for learning in a college-level computer science class sheds light on the crucial relationship between curriculum, instruction, assessment, and self-regulation. Her findings underscore the positive impact of integrating self-regulated learning and assessment for learning into the course, enhancing students' support for the interplay between these elements. By revising the curriculum, instruction, and assessment practices, Chen created a framework that promotes the co-regulation of learning between teachers and students throughout the assessment process. Furthermore, the study emphasizes the need for teachers to actively engage in the co-regulation of learning with their students. This engagement involves providing guidance, feedback, and support throughout the assessment process, empowering students to become self-regulated learners. Educators can create a more inclusive and supportive classroom environment that caters to diverse learning needs by fostering a collaborative approach to learning and assessment. By aligning assessment practices with the principles of self-regulated learning, educators can promote student success and create a dynamic and inclusive learning environment. This approach empowers students to become independent and self-regulated learners and helps educators become self-regulated learners.

Artzt and her colleagues (Artzt et al., 2015) devised a comprehensive model to assess reflective practices among pre-service mathematics teachers. This model consists of three distinct stages corresponding to Zimmerman's three phases of self-regulation. The initial, proactive stage involves teachers engaging in meticulous planning for learning and preparing to deliver their lessons. The interactive stage requires teachers to monitor and regulate the learning process while continually assessing and modifying their actions based on the efficacy of the progress. During this time, teachers are tasked with anticipating questions and reactions from students, all the while actively eliciting participation from their students. Finally, the postactive stage requires teachers to self-evaluate and revise their lessons and class activities based on their self-reflection, subsequently adapting their approach accordingly.

Researchers have successfully implemented Artzt et al.'s (2015) model. For instance, Artzt and Armour-Thomas (1999) reported that teachers who prioritize the development of students' understanding and incorporate instructional strategies into their curriculum and instruction are responsive and self-reflective about their teaching methods and assessments. This approach fosters a proactive learning environment for students. Educators can effectively build a solid foundation for their students' learning journey by integrating such instructional strategies into their teaching practice. This integration aligns with the notion that proactive learners are more likely to take ownership of their learning process, enhancing their educational experience. By providing a structured model that aligns with the phases of self-regulation, these researchers have empowered educators to cultivate reflective teaching practices, thereby enhancing the quality of education for students. Additionally, the emphasis on incorporating instructional strategies and fostering a proactive learning environment underscores the pivotal role of teachers in shaping students' learning experiences. As such, the impact of this work extends beyond individual teachers to benefit the broader educational landscape through effective assessment.

My recent modification to integrating curriculum, instruction, and assessment; which incorporated a cyclical self-regulated learning process, has proven to be highly effective in facilitating the understanding and application of learning theories among teacher candidates. By integrating Bandura's social cognitive theory, Piaget's cognitive developmental theory, and Vygotsky's sociocultural theory in a self-regulated manner, students could engage in a structured approach to mastering these theories. Incorporating self-assessments, such as self-monitoring during the writing process, allowed the students and me to assess their self-efficacy, interests, strategies, and goals before commencing their writing, enhancing their forethought phase. Furthermore, inviting students to reflect and assess their self-efficacy, delay of gratification, help-seeking, and self-monitoring during the performance phase provided valuable insights into their writing process. The self-reflection phase at the end of the writing time enabled students to evaluate their performance, express their self-satisfaction, and assess outcomes and feedback. Implementing this cyclical self-regulated learning process resulted in a high level of motivation among students, as evidenced by their exit ticket responses, and significantly improved grades in their written assignments.

Students also transferred the cyclical self-regulated learning process to other college classes and student teaching with their students. Moreover, the successful application of the cyclical self-regulated learning process has extended beyond the classroom, with students reporting they transferred these valuable skills to other college classes and during their student teaching experiences. This transferability underscores the enduring impact of this approach on students' learning and professional development. The positive outcomes observed in student satisfaction and academic performance highlight the effectiveness of integrating self-regulated learning strategies within the curriculum. As such, this pedagogical approach fosters a deep understanding of learning theories and equips teacher candidates with essential skills they can apply in their future teaching practices. Overall, incorporating a cyclical self-regulated learning process has proven to be a valuable addition to the curriculum and assessment, fostering meaningful learning experiences and empowering students to become self-regulated learners with a heightened sense of efficacy and adaptability.

Educational Implications

Framing curriculum, instruction, and assessment from the perspectives of self-regulated learning highlights four significant hallmarks. By integrating these hallmarks into teaching practices, educators can create a more student-centered and engaging learning environment that reassures students with feedback guidance, encourages them to take accountability for their learning, and develops lifelong learning skills.

First, the iterative position of self-regulated learning emphasizes the learners as agentic individuals capable of proactive and self-directed learning in pursuing academic goals. Learners are also capable of self-assessment and self-reflection of learning outcomes. Similarly, teachers are construed as agentic self-regulated educators in control of their curriculum, instruction, and assessment. Teachers and learners engage in self-regulation, socially shared regulation, and co-regulated learning. As outlined by Greene, Bernacki, and Hadwin (2024) and Hadwin, Järvelä, and Miller (2018), students can be self-regulated learners. Teachers can also be self-regulated learners competent in enactive forethought, self-monitoring, and self-reflection. Students and teachers can work together to create a more effective and engaging learning environment by engaging in self-regulated learning

practices. This approach to education encourages a collaborative and supportive learning community where learners and teachers support each other in pursuing academic goals.

The second hallmark is the adoption of culturally self-regulated pedagogy, an essential focus for educators (Bembenutty, 2023; White & Bembenutty, 2014, 2016). Culturally self-regulated pedagogy emphasizes creating an educational assessment system that is not only diverse and equitable, but also deeply inclusive. By embracing this approach, educators can create an educational system that values and respects all students and teachers regardless of their background or circumstances. Integrating self-regulated learning into teaching practices can help create a better learning environment for all. By focusing on student agency and control, metacognitive and reflective practices, and the role of feedback and self-evaluation, educators can help students develop lifelong learning skills. Furthermore, achieving outcomes beyond successful performance and achievement and embracing a culturally self-regulated pedagogy can help create a more diverse, equitable, and inclusive educational system that benefits everyone.

The third hallmark is self-efficacy, associated with perseverance, persistence, self-control, academic delay of gratification, effort, and emotion regulation. Self-efficacy for teaching relates to teachers' effective classroom management, planning, and imparting effective instruction and assessment. The culturally self-regulated pedagogy model conceives self-efficacy as a foundation for valid and reliable curriculum, instruction, and assessment. The efficacy belief is not a global or personality trait within this dynamic pedagogy. Instead, it is a belief system that operates according to factors structured in the environment, the person, and the behavior (Bandura, 1997).

The fourth hallmark highlights the culturally self-regulated pedagogy's adoption of the principles for assessment in the service of learning (Armour-Thomas & Gordon, 2013; Baker et al., this volume). Specifically, this model endorses the principle that assessment transparency assists teachers, learners, administrators, and parents in understanding learning outcomes. Another principle is that effective assessment results in positive academic outcomes for students' self-regulated learning and can enhance teachers' ability to adopt effective curriculum, instruction, and assessment. This model can benefit learners and educators from diverse

backgrounds who aspire to learn, teach, and assess in inclusive and equitable classroom environments (Armour-Thomas & Gordon, 2013).

Another essential principle ingrained in this model is that Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences. The emphasis on assessment for positive academic outcomes and practical curriculum, instruction, and assessment can benefit learners and teachers from diverse backgrounds. In this sense, teachers are responsible for engaging learners in learning through equitable and fair assessment that can promote and celebrate equity and diversity while instructing and assessing student learning (White & Bembenutty, 2014). This model reflects an equitable educational assessment system in which self-regulated learning facilitates curriculum, instruction, and assessment that can benefit both learners and educators.

Future Research Directions and Conclusion

Beyond just successful performance and achievement, effective curriculum, instruction, and assessment outcomes should include rigor, love, freedom, and joy as outcomes assessments for students and teachers beyond just successful performance and achievement (Zusho et al., 2024). Embracing a new paradigm of standards-based reform can help transform outcomes to achieve these goals. It requires a transformation in the vision and implementation of curriculum, instruction, and assessment. Future research should explore how these four outcomes influence the curriculum, instruction, and assessment in reciprocal interactions (Bandura, 1997).

This deliberate integration of self-regulated learning principles into both in-class activities and homework assignments demonstrates a commitment to fostering students' self-directed learning skills. By engaging in a cyclical process of goal setting, monitoring, and reflection, students are empowered to take ownership of their learning and develop crucial metacognitive abilities. The teacher's role as a facilitator of this process further reinforces the importance of self-regulated learning within the classroom environment. A dynamic assessment system holds promise for cultivating lifelong learners adept at setting goals, monitoring their progress, and reflecting on their learning experiences.

The proactive implementation of the culturally self-regulated dynamic pedagogy assessment model in traditional classrooms is a significant step towards addressing the multifaceted challenges presented by the student population. By integrating self-assessment practices (León et al., 2023) and peer-assessment (Panadero et al., 2023) into the instructional strategies, the educators aimed to cultivate a culture of reflection and self-awareness among the students. This, in turn, was envisioned to contribute towards enhancing their metacognitive skills and fostering a deeper understanding of their learning processes. Furthermore, the emphasis on self-regulated learning aligns with contemporary educational paradigms that recognize the significance of nurturing students' ability to monitor, regulate, and adapt their learning strategies. In doing so, educators are sought to equip students with essential skills for lifelong learning and academic success, transcending the immediate challenges they may face.

Implementing the culturally self-regulated dynamic pedagogy assessment model in an environment characterized by diverse academic abilities and behavioral issues represents a significant step toward promoting inclusive and personalized learning experiences. By foregrounding students' agency in their educational journey, this approach not only addresses immediate academic needs, but also contributes to the holistic development of the students, empowering them to become self-regulated learners capable of navigating complex educational landscapes. However, students need to be ingrained in an educational learning environment that endorses a dynamic system of assessment. The teacher's adaptation of the curriculum, assessment, and instructional approach to incorporate self-regulated learning significantly promotes student autonomy and metacognitive skills.

Conclusion

This chapter underscores the importance of considering the interconnected impact of curriculum, instruction, and assessment on the overall educational experience. It emphasizes the potential for these components to influence the learning journey for both students and educators profoundly. I share the perspective of Armour-Thomas and Gordon (2013) in advocating for the "functional integration of assessment, curriculum, and instruction as instrumental to learning and as the essential components of pedagogy" (p. 2). Their argument for assessment that proactively contributes to student improvement, along with their conceptualization of Dynamic Pedagogy as a pivotal element, has deeply influenced my approach to teaching, self-assessment, and student assessment. I am deeply appreciative of their significant contributions and their role in shaping my professional outlook.

The Culturally Self-Regulated Dynamic Pedagogy Assessment System builds upon the model proposed by Armour-Thomas and Gordon (2013) by emphasizing the significance of a culturally attuned and self-regulated curriculum, instruction, and assessment within our educational framework. This approach aims to elevate the affordances and address the constraints of both learners and educators, leading to positive outcomes for all involved. This chapter encourages readers to recognize that self-regulated learning and cultural considerations are paramount in curriculum, instruction, and assessment. In a dynamic pedagogy assessment system, self-regulated learning and culture matter.

References

Allen, B. A., & Armour-Thomas, E. (1993). Construct validation of metacognition. *The Journal of Psychology, 127*(2), 203–211. https://doi.org/10.1080/00223980.1993.9915555

Armour-Thomas, E. (2008). In search of evidence for the effectiveness of professional development: An exploratory study. *Journal of Urban Learning, Teaching, and Research, 4*, 1–12.

Armour-Thomas, E. (2017). The special role of schooling in the development of academic ability of children and youth. In E. Gordon, B. Jean-Louis, & N. Obiora (Eds.), *Strengthening Families, Communities, and Schools to Support Children's Development* (pp. 63–81). Routledge.

Armour-Thomas, E., & Gordon, E. W. (2013). *Toward an understanding of assessment as a dynamic component of pedagogy.* Educational Testing Service.

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.

Artzt, A. F., & Armour-Thomas, E. (1999). A cognitive model for examining teachers' instructional practice in mathematics: A guide for facilitating teacher reflection. *Educational Studies in Mathematics, 40*(3), 211–235.

Artzt, A. F., Armour-Thomas, E., Curcio, F. R., & Gurl, T. J. (2015). *Becoming a reflective mathematics teacher: A guide for observations and self-assessment.* Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science, 1*(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Bembenutty, H. (2023). Self-regulated learning with computer-based learning environments. *New Directions for Teaching and Learning, 174,* 11–15. Wiley. https://doi.org/10.1002/tl.20543

Bembenutty, H., & Hayes, A. (2018). The triumph of homework completion: Instructional approaches promoting self-regulation of learning and performance among high school learners. In M. K. DiBenedetto (Ed.), *Connecting self-regulated learning and performance with instruction across high school content areas* (pp. 443–470). Springer.

Bembenutty, H., Kitsantas, A., & Cleary, T. J. (Eds.). (2013). *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman*. Information Age Publishing.

Bembenutty, H., Liem, G. A. D., Allen, K.-A., King, R. B., Martin, A. J., Marsh, H. W., Craven, R. G., Kaplan, A., Schunk, D. H., DiBenedetto, M. K., & Datu, J. A. D. (2023). Culture, motivation, self-regulation, and the impactful work of Dennis M. McInerney. *Educational Psychology Review, 35*(1), 28. https://doi.org/10.1007/s10648-023-09743-3

Bembenutty, H., White, M. C., & Vélez, M. R. (2015). *Developing self-regulation of learning and teaching skills among teacher candidates*. Springer.

Bondie, R., & Zusho, A. (2018). *Differentiated instruction made practical: Engaging the extremes through classroom routines*. Routledge.

Bonner, S., & Chen, P. P. (2019). *Systematic classroom assessment: An approach for learning and self-regulation*. Routledge. https://doi.org/10.4324/9781315123127

Butler, D. L., Schnellert, L., & Perry, N. E. (2017). *Developing self-regulating learners*. Pearson.

Chen, P. P. (2023). Interactions between self-regulated learning and assessment for learning in an undergraduate introductory computer science course. *New Directions for Teaching and Learning, 174,* 49–56. https://doi.org/10.1002/tl.20548

Chen, P. P., & Bonner, S. M. (2020). A framework for classroom assessment, learning, and self-regulation, *Assessment in Education: Principles, Policy & Practice, 27*(4), 373–393. https://doi.org/10.1080/0969594X.2019.1619515

Chen, P. P., & Bonner, S. M. (2023). Teachers' Beliefs About Grading, Grades, and Student Classroom Conduct. *Educational Practice and Theory, 45*(2), 69–91. James Nicholas Publishers. https://doi.org/10.7459/ept/45.2.06

Cleary, T. J., & Russo, M. R. (2024). A multilevel framework for assessing self-regulated learning in school contexts: Innovations, challenges, and future directions. *Psychology in the Schools, 61*(1), 80–102. https://doi.org/10.1002/pits.23035

DiBenedetto, M. K. (Ed.).(2018). *Connecting self-regulated learning and performance with instruction across high school content areas.* Springer.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20,* 391–409. (2018). *Culturally responsive teaching: Theory, research, and practice.* Teachers College Press.

Golmohammadi, L. (2022). *How to use online probes for social science research.* SAGE.

Greene, J. A., Bernacki, M. L., & Hadwin, A. F. (2024). Self-regulation. In P. A. Schutz & K. R. Muis (Eds.), *Handbook of educational psychology* (pp. 314–334). Routledge. https://doi.org/10.4324/9780429433726-17

Hacker, D. J., & Dunlosky, J. (2003). Not all metacognition is created equal. *New Directions for Teaching & Learning, 95*, 73–79. https://doi.org/10.1002/tl.116

Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 83–106). Routledge.

Hoy, A. W., Hoy, W. K., & Davis, H. A. (2009). Teachers' Self-Efficacy Beliefs. *In Handbook of motivation at school* (pp. 641–668). Routledge.

Jenkins, J. L., & Shoopman, B. T. (2019). Identifying misconceptions that limit student understanding of molecular orbital diagrams. *Science Education International, 30*(3), 152–157. https://doi.org/10.33828/sei.v30.i3.1

Kaplan, A., Neuber, A., & Garner, J. K. (2019). An identity systems perspective on high ability in self-regulated learning. *High Ability Studies, 30*(1–2), 53–78. https://doi.org/10.1080/13598139.2019.1568830

Karabenick, S. A., & Gonida, E. N. (2018). Academic help seeking as a self-regulated learning strategy: Current issues, future directions. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 421–433). Routledge.

Kitsantas, A., Cleary, T. J., DiBenedetto, M. K., & Hiller, S. E. (2024). *Essentials of research methods for educators*. SAGE.

Kramarski, B., & Kohen, Z. (2017). Promoting preservice teachers' dual self-regulation roles as learners and as teachers: Effects of generic vs. specific prompts. *Metacognition and Learning, 12,* 157–191. https://doi.org/10.1007/s11409-016-9164-8

Ladson-Billings, G. (2021). *Culturally relevant pedagogy: Asking a different question.* Teachers College Press.

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review, 35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0

Matthews, J. S., & Wigfield, A. (2024). Past due! Racializing aspects of situated expectancy-value theory through the lens of critical race theory. *Motivation Science*. Advanced online publication. https://doi.org/10.1037/mot0000337

Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education, 48*(8), 1053–1075. https://doi.org/10.1080/02602938.2023.2164884

Pape, S. J., Bell, C. V., & Yetkin-Ozdemir, I. E. (2013). Sequencing components of mathematics lessons to maximize development of self-regulation: Theory, practice, and intervention. In H. Bembenutty, T. J. Cleary, & A. Kitsantas (Eds.), *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (pp. 29–58). Information Age Publishing.

Schunk, D. H., & Greene, J. A. (Eds.). (2018).*Handbook of self-regulation of learning and performance*. Routledge.

Slavin, R. E. (2018). *Educational psychology: Theory and practice*. Pearson.

Täschner, J., Dicke, T., Reinhold, S., & Holzberger, D. (2024). "Yes, I Can!" A systematic review and meta-analysis of intervention studies promoting teacher self-efficacy. *Review of Educational Research*. https://doi.org/10.3102/00346543231221499

Veenman, M. V. J. (2007). The assessment and instruction of self-regulation in computer-based environments: a discussion. *Metacognition and Learning, 2,* 177–183. https://doi.org/10.1007/s11409-007-9017-6

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning, 6*, 205–211. https://doi.org/10.1007/s11409-011-9080-x

White, M. C. (2017). Cognitive modeling and self-regulation of learning in instructional settings. *Teachers College Record, 119*, 1–26.

White, M. C., & Bembenutty, H. (2014, October). *Teachers as culturally proactive agents through cycles of self-regulation.* Paper presented at the Dept. of Secondary Education and Youth Services Research Symposium, Queens College, New York, NY.

White, M. C., & DiBenedetto, M. K. (2015). *Self-regulation and the common core: Application to ELA standards*. Routledge.

White, M. C., & DiBenedetto, M. K. (2018). Self-regulation: An integral part of standards-based education. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 208–222). Routledge. https://doi.org/10.4324/9781315697048-14

Woolfolk Hoy, A., & Weinstein, C. S. (2006). Students' and teachers' perspectives on classroom management. In C. Evertson, & C. S. Weinstein (Eds.), *Handbook for classroom management: Research, practice, and contemporary issues* (pp. 181–220). Erlbaum.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25,* 82–91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist, 48*(3), 135–147. https://doi.org/10.1080/00461520.2013.794676

Appendix

Figure 1.

Barry J. Zimmerman's cyclical model of self-regulated learning.

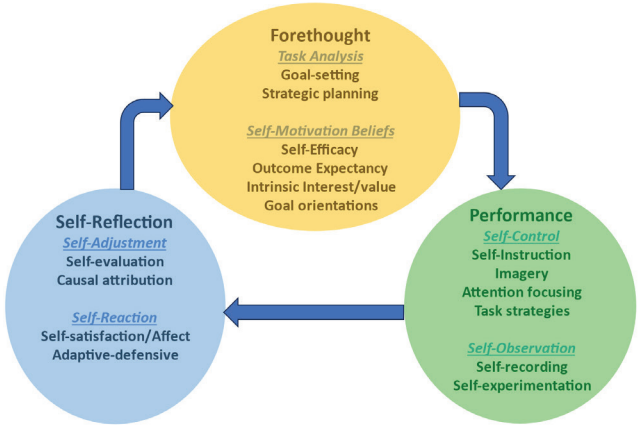


Figure 2.

Armour-Thomas and Gordon's Dynamic Pedagogy Framework

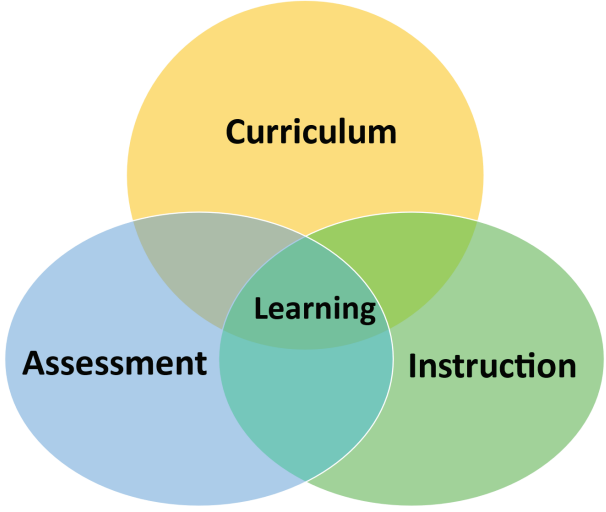


Figure 3.
Culturally Self-Regulated Dynamic Pedagogy Assessment Model

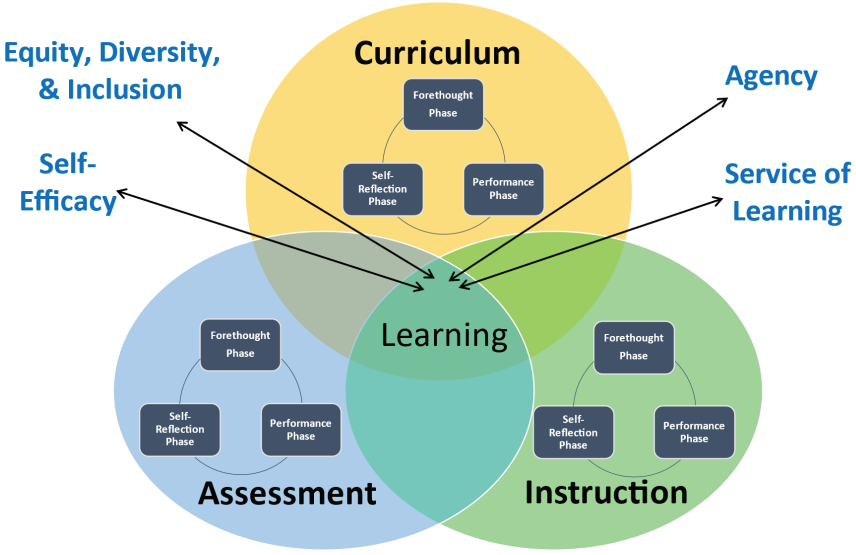


Table 1

Developing Educational Assessments to Serve Learners

Assessment Dynamic Pedagogy	Culturally Self-Regulated Dynamic Pedagogy Assessment System
Curriculum Strand	
• Encompasses the fundamental ideas, rules, criteria, and resources that teachers utilize to facilitate learning effectively and involves the careful arrangement and communication of content knowledge in a manner accessible to learners. • Aims to deliver content at a suitable level for the target audience, ensuring that it is tailored appropriately to their needs and abilities. • Aligns closely with the evaluation of students' learning outcomes ensuring assessment is focused, fair, and accurate.	• Embeds self-regulated learning within its structure, encompassing three cyclical phases: forethought, performance, and self-reflection. • Guides by the principles of cultural self-regulated pedagogy to foster self-regulated learning among diverse learners. • Provides clear learning goals, expectations, strategies, and resources to facilitate effective planning and self-motivation in the forethought phase. • Offers a diverse range of media and formats to deliver content, catering to various learning preferences and styles, and students are provided with opportunities to seek feedback and monitor their progress in the performance phase. • Incorporates tools and activities that enable students to evaluate their learning outcomes, including their attributions and adaptability in the self-reflection phase.

Continued on the next page

Table 1. (continued)

Assessment Dynamic Pedagogy	Culturally Self-Regulated Dynamic Pedagogy Assessment System
Instruction Strand	
• Consists of strategies helpful to facilitate learning, including guided practice, supervised independent practice, modeling, scaffolding, and peer learning. • Relates to assessment by revealing strengths and limitations in the assessment process. • Facilitates assessment feedback, which teachers can use to implement new instructional approaches that could result in effective learning.	• Embeds self-regulation within the instruction dimension, including three phases shaping the self-regulated learning processes. • Creates opportunities for self-assessing self-efficacy, interest, and task value. Teachers can model ways to set measurable and realistic goals and identify learning goals and strategies in the forethought phase. • Helps self-monitor progress by providing self-monitoring forms or logs and inviting them to seek help and assess their self-efficacy again during the performance phase. • Engages students in self-assessment or practice peer assessment and self-evaluation and helps them adopt appropriate attributions for academic success or failure in the self-reflection phase. • Adopts a culturally self-regulated pedagogy in the classroom, and teachers and students are construed as agents willing to engage in socially shared regulation and co-regulation.

Continued on the next page

Table 1. (continued)

Assessment Dynamic Pedagogy	Culturally Self-Regulated Dynamic Pedagogy Assessment System
Assessment Strand	
• Consists of strategies helpful to facilitate learning, guided practice, independent practice, modeling, scaffolding, and peer learning. • Relates to assessment by revealing strengths and limitations in the assessment process. • Implements instruction that could result in effective learning given assessment feedback. • Includes two probes. The online probe helps teachers assess students' prior knowledge, skills, and readiness for new learning, while the metacognitive probe helps students become aware of effective learning strategies.	• Embeds self-regulation within the assessment dimension, including three phases shaping the self-regulated learning processes. • Embraces assessment that is culturally sensitive, validated, and reliable. • Ensures the assessment has undergone a rigorous task analysis, activated prior knowledge, and enabled students to use strategies within reasonable self-efficacy beliefs in the forethought phase. • Allows students to successfully apply and monitor goals and strategies to complete the tasks in the performance phase. • Serves as a tool for self-evaluation, provides feedback, and conveys expectations that learning is possible in the self-reflection phase. • Offers opportunities for diverse ways of responding while it is culturally fair.
Note: The Culturally Self-Regulated Dynamic Pedagogy Assessment System includes all the functions outlined in the Assessment Dynamic Pedagogy model.	

Table 2

Comparing Curriculum between Culturally Self-Regulated Pedagogy (CSP) and Culturally Responsive Teaching (CRT)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Content Engagement			
Design content that includes cultural values for practicing self-regulation, such as research projects.	Design content that reflects students' cultural backgrounds, making learning more relevant and meaningful.	Engage with content that includes activities for practicing self-regulation.	Engage with content that reflects their cultural backgrounds, making learning more relevant and meaningful.
Curriculum Goals			
Set curriculum goals that encourage students to develop self-regulation skills and cultural awareness.	Set curriculum goals that foster cultural awareness and respect diverse cultural backgrounds.	Set goals and develop plans to achieve them, focusing on self-regulation skills.	See their cultural identities reflected in the curriculum goals and a sense of belonging and relevance.
Resource Utilization			
Provide resources (e.g., self-monitoring forms, homework logs) to support students' self-regulated learning.	Provide culturally diverse resources that reflect students' cultural backgrounds and experiences.	Use resources like planners and goal-setting templates to support their self-regulated learning.	Access culturally diverse resources that reflect their cultural backgrounds and experiences.
Curriculum Relevance			
Select topics for research projects that align with students' personal interests and academic goals, fostering self-regulation.	Choose research topics that reflect students' cultural backgrounds and experiences, making learning more meaningful and engaging.	Select topics for research projects that align with their personal interests and academic goals, fostering self-regulation.	Choose research topics that reflect their cultural backgrounds and experiences, making learning more meaningful and engaging.

Continued on the next page

Table 2. (continued)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Technology Integration			
Integrate technology tools that support cultural diversity and self-regulated learning, such as goal setting and progress-tracking apps.	Use technology to provide access to culturally diverse resources and understanding of diverse cultures.	Use technology tools that support self-regulated learning, such as goal setting and progress-tracking apps.	Use technology to access culturally diverse resources and materials, enhancing their understanding of diverse cultures.
Independent Learning			
Design independent learning activities that require students to set goals, plan their work, and monitor their progress within their cultural interests.	Design independent learning activities incorporating students' cultural interests and experiences, making learning more engaging	Engage in independent and proactive learning activities that require them to set goals, plan, and monitor their progress.	Participate in independent learning activities incorporating their cultural interests and experiences, making learning more engaging.

Table 3

Comparing Instruction between Culturally Self-Regulated Pedagogy (CSP) and Culturally Responsive Teaching (CRT)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Learning Strategies			
Use strategies that promote self-regulated learning, such as teaching students how to set goals, monitor their progress, and adjust their strategy.	Employ culturally responsive instructional strategies that reflect students' cultural identities and experiences, making learning more relatable.	Learn and apply self-regulation strategies, such as goal setting, time management, and self-assessment.	Participate in culturally responsive learning activities that incorporate their cultural experiences and perspectives.
Student Autonomy and Peer Feedback			
Encourage autonomy by allowing students to choose their learning activities and set goals and encourage students to provide and receive peer feedback on their self-regulation strategies,	Incorporate students' cultural practices and preferences into the learning process, allowing culturally relevant choices in learning activities and facilitating culturally sensitive peer feedback	Take ownership of their learning by setting their own goals, monitoring their progress and providing and receiving peer feedback on their self-regulation strategies,	Have the opportunity to make culturally relevant choices in their learning activities, enhancing engagement and motivation and give and receive culturally sensitive peer feedback.
Independent and Collaborative Learning			
Promote independent and collaborative learning activities that help students develop self-regulation skills, where students set goals and monitor their progress.	Facilitate independent and collaborative learning activities that encourage cultural exchange, allowing students to learn from each other's diverse cultural perspectives.	Work independently and collaborate with peers to set group goals and monitor progress, developing self-regulation skills.	Engage in independent and collaborative learning activities that promote cultural exchange and understanding.

Continued on the next page

Table 3. (continued)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Diverse Instruction			
Use diverse instruction to cater to students' self-regulation needs, providing various support and resources based on students' self-regulation skills.	Use differentiated instruction to address diverse cultural backgrounds, ensuring all students can access culturally relevant experiences.	Receive diverse instruction based on their individual self-regulation needs, with varying levels of support.	Benefit from differentiated instruction that addresses their diverse cultural backgrounds, ensuring meaningful learning experiences.
Motivation and Self-Efficacy			
Use motivational approaches that promote self-efficacy and self-regulation, such as setting and rewarding incremental goals.	Use culturally relevant motivation techniques to increase engagement, such as incorporating students' cultural interests and values into the learning process.	Use motivation approaches that promote self-regulation (e.g., setting incremental goals, providing rewards and self-efficacy).	Use culturally relevant motivation techniques, such as incorporating their cultural interests and values into the learning process to increase engagement.

Table 4

Comparing Assessment between Culturally Self-Regulated Pedagogy (CSP) and Culturally Responsive Teaching (CRT)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Self-Assessment			
Use self-assessment tools to help students reflect on their learning and identify areas for improvement and content knowledge.	Use culturally responsive assessments that consider students' cultural backgrounds and understanding.	Use self-assessment tools to reflect on their learning and identify areas for improvement.	Participate in culturally responsive assessments that consider their cultural backgrounds and understanding.
Formative Assessment			
Provide formative feedback that helps students develop self-regulation skills and adjust their learning strategies.	Give culturally sensitive feedback that acknowledges students' cultural identities and supports their academic growth.	Receive formative feedback that helps them develop self-regulation and content skills and adjust their learning strategies.	Receive culturally sensitive feedback that acknowledges their cultural identities and supports their academic growth.
Summative Assessment			
Design summative assessments that require students to demonstrate their self-regulation skills and content knowledge, such as comprehensive projects or portfolios.	Design summative assessments that allow students to showcase their cultural knowledge through culturally relevant projects.	Complete summative assessments requiring the demonstration of self-regulation skills, such as comprehensive projects or portfolios within specific content.	Engage in summative assessments that allow them to showcase their cultural knowledge and perspectives, such as through culturally relevant projects.

Continued on the next page

Table 4. (continued)

CSP	CRT	CSP	CRT
Assessment: TEACHERS		**Assessment: STUDENTS**	
Performance Tasks			
Design performance tasks requiring students to demonstrate self-regulation skills, such as managing a long-term project and identity.	Design performance tasks that allow students to showcase their cultural knowledge (e.g., a cultural presentation).	Complete performance tasks requiring the demonstration of self-regulation skills, such as managing a long-term project.	Engage in performance tasks that allow them to showcase their cultural knowledge, such as creating a cultural presentation.
Self-Reflective Tools			
Encourage students to keep reflective journals and self-monitor tools to track their progress and reflect on their self-regulation strategies and cultural awareness.	Encourage students to use reflective journals to explore their cultural identities and how their cultural experiences influence their learning.	Keep reflective journals and logs to track their progress and reflect on their self-regulation strategies and cultural experiences.	Use reflective journals to explore their cultural identities and how their cultural experiences influence their learning.

Personalizing Assessment for the Advancement of Equity and Learning

Randy E. Bennett, Eva L. Baker, and Edmund W. Gordon

Abstract

The purpose of this chapter is to discuss two of the *Principles for Assessment in the Service of Learning* as they affect the design and use of assessments intended to personalize learning for greater impact and equity:

- Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.
- Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

Our discussion begins with background that gives the context for why these principles are needed. That background is followed by a section on some of the various strategies that might be used to achieve the two principles, as well as related design and application issues. The strategies explored are necessarily aspirational in that the state-of-the-art does not yet support their realization at scale. In the third section, we discuss sources of evidence related to quality. We close the chapter with overarching remarks.

Background

In many countries around the world, dramatic population shifts have resulted in unprecedented levels of heterogeneity. Termed "superdiversity" by Vertovec (2023), this condition has come in significant part from immigration, as well as from such other sources as differential birth rates and intermarriage. Concerning race/ethnicity, for example, the U.S. population has evolved from 80% Caucasian in 1980 to 59% in 2021 (Hobbs & Stoops, 2002; US Census Bureau, 2021), a doubling in the proportion of people of color from approximately 20% to 40%. As of 2018, the U.S. public-school population was 53% students of color (National Center for Education Statistics, 2020, Tables 203.70). Throughout Europe, as well as in other parts of the world, countries are experiencing similarly dramatic demographic shifts (Vertovec, 2023).

Many human characteristics are involved in these demographic shifts. Although the most visible might be race/ethnicity, heterogeneity in many locales is also growing along other dimensions. These dimensions include socioeconomic class, immigrant status, native language, gender identity, exceptionality, and sexual orientation. Greater levels of cultural heterogeneity inevitably arise as a consequence of these various kinds of diversity.

Cultural diversity, however, poses a significant challenge (and significant opportunities) for educational assessment in that assessments are cultural artifacts (Greenfield, 1997; Solano-Flores, 2019). Those artifacts presume certain modes of expression, ways of knowing, linguistic structures, and forms of representation. Although populations have changed dramatically, our assessments have remained relatively fixed over time, creating a misalignment. To illustrate, the most obvious change in operational testing programs over the same 1980 to 2021 period in which the US experienced a dramatic shift in racial/ethnic composition was the transition from paper to computer delivery, not a very deep change at all. Classroom assessment practice also has not kept pace in that the demographic composition of public-school teachers is very different from that of students. In general, relatively little investment appears to have been made by federal or state governments in helping a largely Caucasian and middle-class teaching force adjust their assessment and instructional methodologies to the high levels of cultural diversity with which many teachers are confronted.

Noting this growing need, various proposals have been made for accommodating variation in learner populations. Those approaches include culturally responsive assessment (Hood, 1998; Lee, 1998), justice-oriented assessment (Randall, 2021), socioculturally responsive assessment (Bennett, 2023), antiracist assessment (Inoue, 2015), culturally sustaining classroom assessment (Lyons et al., 2021), and universal design for assessment (Ketterlin-Geller, 2005). These approaches may vary along such dimensions as definition, goals, principles espoused, populations targeted, underlying literatures, and intended setting (Bennett, 2025). Even so, most such approaches have at least one shared idea: Designing assessment for the social, cultural, and other relevant characteristics of individuals and the contexts from which they come.

Among the many issues these approaches raise is how to conceptualize diversity, as that conceptualization dictates how to respond in an assessment context. Crenshaw's (1989) insight was that individuals might have multiple identities that could lead to multiple marginalizations. In the context of completing government forms, Rogoff (2003; Rogoff & Angelillo, 2002) described the "box" problem as perceiving groups as homogenous and mutually exclusive when, in fact, they are not. Lee and colleagues (Lee et al., 2020) observed that individuals belong to multiple, not single, cultural groups. Building on these ideas, Vertovec (2023) advocated for "category +," the idea that an individual might primarily affiliate with one group, but secondarily identify with several others. Vertovec argued that such multiple affiliations increase tolerance as one creates more heterogenous connections. Gordon et al. (2019) and Darling-Hammond (personal communication, December 19, 2022) push this idea further still by arguing for an individual-differences or human-variance approach, whereby every learner brings a unique sociocultural identity to learning and assessment.

These conceptualizations of diversity, especially those of Gordon et al. (2019) and Darling-Hammond (personal communication, December 19, 2022), argue strongly for adaptation. Adaptation in educational assessment, and in education more generally, is not new. Adjusting to competency level dates at least to the Stanford-Binet Intelligence Scales, in which the examiner adjusted the difficulty of questions asked based on the answers given by the examinee. That idea appeared in the 1980s in a more theoretically grounded form as computerized adaptive testing (Ward, 1988). Separately, real-time, theory-based adaptation to an individual's domain competencies characterized intelligent tutoring systems (Sleeman & Brown,

1982). Both adaptive tests and adaptive learning systems are now widespread. Other types of adaptation also are currently found in operational assessment programs, including adaptation to exceptionality (in every major test), native language (in many state accountability tests and every international group-score assessment), and interests (various Advanced Placement examinations allowing problem choice or even problem design).

The goals of adaptation in these programs and in the aspirational assessments we envision are described in the section on Evidence of Quality. For now, it is sufficient to note that those goals extend to enhancing such factors as identification with the assessment, engagement and motivation to perform, test performance, learning, perceptions of the test as fair, cultural identity, and meaningful interpretation of results. Some of these factors (e.g., engagement and motivation to perform, cultural identity) will be impacted far more significantly by influences outside of the assessment (e.g., family, community). The goal of assessment adaptation is enhancement, however small such enhancement may prove to be relative to other influences.

Strategies for Adaptation

In this section, we consider approaches to adapting learning and assessment to individual differences, including to the needs and desires of learners. Like many good ideas, it is easier said than done.

Central to the consideration of adaptation are the following questions:

1. What are the purposes of adaptation?
2. What is to be learned and/or assessed?
3. What is to be adapted?
4. On what basis does the adaptation occur?
5. How does adaptation interact with complex learning and assessments?
6. What evidence can be developed in support of adaptation?

Adaptation can be conceived of as multiway matches or interactions among student individual difference variables, instruction, and assessment components or measures that react to interactions. In other words, employed adaptations should improve the performance of target students.

There is considerable history and present practice using measures of student prior knowledge or ability as the basis of adaptation. In addition, there is persistent interest in variables such as learning style and student preferences, although their effects are not well supported (DeBruyckere et al., 2015; Hattie, 2023). Indeed, the lack of positive findings may shift our strategies for applying adaptive approaches from those with evidence of effectiveness to those which "do no harm." We will discuss the implications of lack of convincing, positive data in the light of the clear need for rapid progress in adaptation and policies supporting the evaluation of learners. As we step through the process, both pitfalls and opportunities will emerge.

1. What are the purposes of adaptation?

Our starting point is that adaptation, or the provision of differential options or experiences for learners, is intended to increase for learners both the effectiveness and the equity in the learning-assessment process (Buzick, Casabianca, & Gholson, 2023). Fairness and equity used to be framed as providing the exact same instruction and assessments to each student. One can consider the early designs of standardized test administration, including timed performance, as the pursuit of uniformity, the hallmark of fairness. Everyone was to be given the same experience. However, as our learning about variations in cultural and personal experiences changed (See for example, Wenger, 1998), so have our formulations of fairness. Instead of thinking about equity as providing uniform exposure in highly specified conditions, we have more recently focused on how elements of the situation and how the content of the material may be interpreted differently and what can be done to assist different respondents to display their most positive or representative level of expertise. Adaptation in learning situations may involve wide-ranging differences in sequence, repetition, complexity, and feedback. In assessment that is a part of learning, adaptation of timing, resources, and content are more easily included in instructional variations based on sequences of learners' answers, errors, and timing. In a formal assessment setting, which typically is far shorter than the instructional experience, adaptations present greater design challenges.

2. What is to be learned and/or assessed?

Assessments vary in many ways, including their purposes, tasks, domains, and variety. It should be expected that specific features of assessments (e.g., multiple steps, reliance on domain knowledge) will trigger varied student reactions resulting in performance differences. For example, outcome measures differ between and within their designs in terms of expected cognitive demands on the assessment. Methods to classify types of cognition in learning or on examinations are both numerous and varied. A major classification branch on assessment arises from the venerable works of Bloom and his associates (1956) as updated by Anderson and Krathwohl (2001). These approaches examine the assessment and its items to characterize their cognitive demands. Other classification systems focus more directly on the learning process, such as those generated by Gagné and Medsker (1996), Merrill (2009), and Webb (1997). In simplified terms, all these systems subdivide learning processes and outcomes into at least two categories: (a) access and understanding knowledge, and (b) applications of knowledge and other thinking processes (See also Hattie et al., this volume).

A third area of great importance, but limited research, is how knowledge and thinking processes may be applied differently to situations ranging from the familiar (those used in the instructional examples) compared to tasks embedded in previously unencountered contexts or problems. In educational psychology, the movement from familiar to unfamiliar situations in performance is called "transfer" (See Perkins & Salomon, 1992). Clearly, when test situations are unfamiliar to students, such as learners from diverse backgrounds, we are functionally adding the requirement of transfer to their learning objectives, as compared with students who may, because of their background, be already accustomed to the contexts and conditions of the assessment.

Research on transfer does in fact indicate that the presentation of unfamiliar contexts and conditions generally diminishes performance, giving one additional explanation for persistent lower performance of some groups. Even seemingly unimportant differences in assessments, such as the use of synonyms for vocabulary, can have an effect (Abedi & Ewers, 2013). The phenomenon of transfer, then, leads to the premise that students who have less broad, or different experiences, will likely perform less well when they are less accustomed to variations in assessment and learning situations.

How to change this situation is not clear. The willingness to engage in unfamiliar tasks may be encouraged by caregivers or the community (See Gordon, 1984), but by itself such encouragement is unlikely to reduce the difficulty of such tasks for the learner. As Gordon (personal communication, August 1, 2024) has asserted, to be successful minoritized students must learn the intellective habits of mind and curricular content dictated by the dominant group, however unfair that might seem. Advocates of culturally attuned approaches to teaching (Gay, 2002, 2018; Ladson-Billings, 1995; 2021; Moll, Amanti, Neff, & Gonzalez, 1992) argue for using the knowledge and identity brought from home and community as conduits to learning such dominantly valued content. Advocates also argue for broadening that dominantly valued content. Broadening has in fact occurred in some states (Bellamy-Walker, 2022; Mays, 2021; NYSED, n.d.), though in other states the opposite appears to be happening (Cavanaugh, 2023).

Thus, despite the logic of making an assessment more familiar to the learner (again, to avoid unintended transfer load for diverse students), the issue is a complex one. As the underlying logic and transfer research might suggest, the resulting scores may not generalize to performance in other contexts, particularly the contexts valued by the dominant group. Moreover, the evidence base in support of adapting instruction and assessment is neither large nor overwhelmingly positive. Given that many research studies used standardized, general assessments of constructs rather than ones more attuned to the types of transfer that might be expected from what has been specifically emphasized in instruction, a new wave of studies of adaptation is obviously needed and is partially described in a subsequent section.

3. What is to be adapted?

Let us begin with fundamental questions about the premises and methods of developing approaches that adapt assessments to learner needs. We then proceed to address approaches that may offer substantial inroads to the solution of a heretofore intractable problem. Our initial considerations focus on what interventions may be applied, while the second set of concerns attends to how we might go about their implementation.

Looking for interventions that help learners with different measured abilities, needs, or preferences has occupied educational psychologists for decades. For instance, in their landmark study investigating ability-treatment interactions in learning, Cronbach and Snow (1977) concluded that the lack of positive findings

may be partly attributed to the character of dependent or outcome measures used in the study (pp. 170–171). Cronbach and Snow suggested that where studies use well-defined outcomes, performance is likely to be more positively affected by adaptations than when more general measures are used. They cited Atkinson's (1972) micro-adaptation as an example. Their research is among many that have sought to discover reliable interactions among learner background variables, attributes of instruction, and assessment systems that positively affect learner performance.

In both learning and assessment, we assume that sets of learners need different experiences, tasks, cues, and sequences to assist their achieving desired performance. Learning sequences and extended examinations involve content or domain knowledge, situations, and tasks, the last of these including requirements, response formats, and stringency levels of judgment. Various terms have been used to describe adjusting testing experiences to support the respondent. Perhaps the most used has been the term "accommodations," generally supporting performance by relaxing time limits, providing additional cues such as glossaries, or enabling opportunity to review items of content (Abedi & Ewers, 2013).

The target of adaptation can be addressed in two large and somewhat overlapping sets of variables. The first focuses on components in the learners' repertoire that clearly relate to the desired performance. These components may reflect essential subcomponents needed for successful performance. The second type of adaptation target is indirectly related, where the variable serves as a moderator to more directly related behaviors.

Variables for Adaptation

The prior knowledge of learners serves as a frequent way to determine adaptation. This approach is more situation and domain specific than measures of general ability considered by Cronbach and Snow (1977). In Hattie's syntheses of meta-analyses of research (2023), he reports a weighted effect size of .96 for prior ability and intelligence, and an effect size of .73 for prior achievement of learners. These findings suggest the potential efficacy of identifying learner differences that are connected directly to identifiable components of instruction and assessment. In fact, such prior knowledge dominates the field of adaptation, whether presented in computer platforms, as lessons or learning games, or in tests. Employed as a strategy in early instances of programmed instruction (Pressey, 1950), prior

knowledge displayed in initial performance has been at the core of computer adaptive testing (CAT) for decades (Weiss, 2011). CAT operates by zeroing in on where the learner is in a framework of assessment items. It uses initial knowledge to bracket and then confirm performance levels. While widespread, a major function of CAT is to make the testing process more accurate and efficient, rather than to support learner performance.

Other direct measures, perhaps observed during the learning/assessment process in computer learning systems, include speed of response and pattern of errors on the tasks of interest. These measures may be created top-down, as a modification of rule-based design, or inferred from deep learning or other bottom-up processes. Using the fundamental design of intelligent systems, described by Mislevy (2018) and by Shute and others (2015), individual response data reside in a student model which is updated as continued responses are made. To determine the best next action, the system may compare the learner's patterns to either expert models or to the repository of patterns obtained from other students. Based on these analyses, a recommendation for the next item or sequence is generated and implemented. This approach can result in varied numbers of responses elicited from the student and shortening or lengthening the learning/assessment experience as needed.

The difficult fact is that students who have been found to lack elements essential for further learning, for instance, mastery of a particular math procedure, will need to acquire it in order to undertake the desired task. Catching up without taking more time is a difficult proposition. One can reflect on approaches that may allow students to overcome missed skills or knowledge without spending considerable portions of available time, where such time is relatively fixed. Certainly, familial motivation is an important factor, but not one that is easily amenable to change (Gordon, personal communication, 2024). One strategy that received strong moral support was advocated by Oakes (1992), who wished to "detrack" learning by placing students in heterogeneous groups in classrooms. Overall effects for detracking are small (ES = .09) and comparable for findings in tracked classroom (ES = .09) (Hattie, 2023). In tracked environments, lowest level students did the worst, perhaps because of the catching-up process. Conversely, in detracked studies, it was lower performing students who were most positively affected.

Most intriguing, perhaps, are findings about approaches intended to increase the student's agency in learning, such as self-judgment and reflection (ES = .69), metacognitive strategies (ES = .52), and self-regulation (ES = .52; Hattie, 2023, p. 349). Brummelman and Sedikides (2023) also argued for approaches to animate students' agency. Nonetheless, student control over instruction does not rank high on the meta-analytic summaries of effects of learning (ES = .04), but ranks somewhat higher in meta-analyses of motivation (ES=.30; Hattie, 2009, p. 193). We can posit a few reasons. First, it may be possible that the options offered are not functionally different. The option to write about "how I spend my free time" versus "my favorite summer vacation" could appear to be identical to a given learner. Developing choices that connect specifically to varied aspects of different learner's experiences is a non-trivial task. A second concern is that even with reasonably different choices, the learner may not know how to select the best personal choice. In this case, the teacher or instructional system should be enabled to give useful (validated) instruction on how to make choices. This instruction might include encouraging the learner to answer a series of questions about the experience with the choice, the need for further information or support, and their ability to recognize a good choice. Such instruction may well be conducted collaboratively.

Beyond the ability to choose procedures or options most likely to result in improved learning, at some point, choice should also include goals and objectives that supplement or depart from the uniform curriculum.

Indirect Variables for Adaptation

Indirect approaches to adaptation focus on variables or experiences that surround and may interact with the learning-assessment process. Among the many examples here are efforts to reduce test anxiety, a variable that has a reliable negative relationship to performance (Rana & Mahmoud, 2010). Hattie describes a .30 effect size on achievement in meta-analyses of test anxiety reduction (Hattie, 2009, pp. 49–50) by providing respondents with choices to control the assessment environment. Other indirect approaches operate on presumably improving motivation through use of engaging or amusing episodes, pictures, or content that is likely to interest the learner. Mayer (2014) and others before him have cautioned that such options, unless carefully designed, can distract the learner, and reduce their likelihood of demonstrating their best performance. Moreover, options that take more time to complete can reduce performance when time limits are imposed.

Another major area of interest is motivation as an intervening variable affecting performance. Bandura (1977) noted positive emotional and physiological states support learning. In 1982, he added comfort and experience in the cultural setting, which can assist learners to perform, a point amplified by the work of Duran (2022). Motivation may be considered as arousal or seeking to engage in learning. Technological approaches to measuring motivation and arousal as processes, using measures of eye movement, pupil dilation, and other sensor-based approaches, have been well summarized by Plass and Kalyuga (2019).

There are also conceptual and practical questions related to these approaches as strategies for equity. A principal concern is that many of these variables often show main effects on learning and performance; that is, they help everyone to some degree. If equity is framed as equal opportunity, this concern should not be central. But in discussions underlying approaches to diverse students, an assumption is often made about differential and potentially accelerated progress, in which case main effects are not as desirable as interactions (Sireci, Scarpati, & Li, 2005).

4. On what basis does the adaptation occur?

Deciding on who gets adaptation may depend on performance on a pre-instructional or initial assessment indicator (or indicators) obtained prior to encountering the bulk of the learning or assessment. In many assessment systems, algorithms are used to capture and modify the learner's progress in the episode of assessment and may well include earlier patterns of performance. Different learners may receive adaptations on different schedules. It is less clear whether there are empirically verifiable cut scores for the decision(s) to adapt performance.

This discussion raises a significant conundrum. How do we avoid stereotyping students based on assumptions about their cultural experiences, language status, and socioeconomic characteristics? Without an accurate picture of the learner, how do we maintain our interest in, and adaptation to, the individual differences the student has, specifically those that may be independent of, or diverge from, those differences generally ascribed to the population group? Consider if the construction of a model of genomic analysis of learning, somewhat like the LEARNOME analogy by Baker (Baker et al., 2002) or that of MIT (2024), might help. Here detailed attributes of, and progress in, the learner's performance and other measured skills are integrated in a description of capability. The complexity of personalization in such models may result in interventions that assist students beyond the use

of group stereotypes. We can make progress on effective adaptation using our present approaches, but serious attainment of equity is not likely without identifying the multitude of critical knowledge, skills, and attributes of students, as well as their interactions.

5. What about complex assessments?

To address the question of adaptation for learning more complex behaviors, and for assessment that embeds learning (e.g., dynamic assessments), we start with the literature in learning. Most stand-alone assessment situations may be of relatively short duration. Students may also recognize the stakes pertaining to their performance. In complex assessments, with an extended process or product, like an oral science report on a set of experiments, an essay based on research, or a process, such as a musical performance, the student manages several individual and interacting elements to complete the task. In the conduct of such complex assessments, support may be given to assure that required prerequisites (such as knowing the right equation, identifying a tractable problem, or accessing suitable examples) are provided. That support may either take the form of additional resources or assets provided during the assessment, or feedback given on adequacy of the learner's formulation of a needed step. Clearly, there is a blurring of the boundaries between assessment and learning, the integration of which for many observers is not a problem. On the other hand, if the assessment is formulated so students are intended to display competence on a generally comparable set of requirements, and if the claimed comparability is the basis of high stakes decisions, then any adaptation that substantially compromises comparability may be out of the question. Rather, we propose thinking about the adjustments here as tuning the assessment to the learner, at once a more precise and less burdensome approach. One option that may meet these desires is giving choices to the learner of content, emphases, and presentation, despite the ambiguous data in its support (Powers & Bennett, 1999).

We have reviewed steps important to the determination of approaches to adaptation including why, what, and how. In the next section we consider the questions that might be used to frame our thinking theoretically about the evidence of adaptation quality (question 6, discussed earlier).

Evidence of Quality

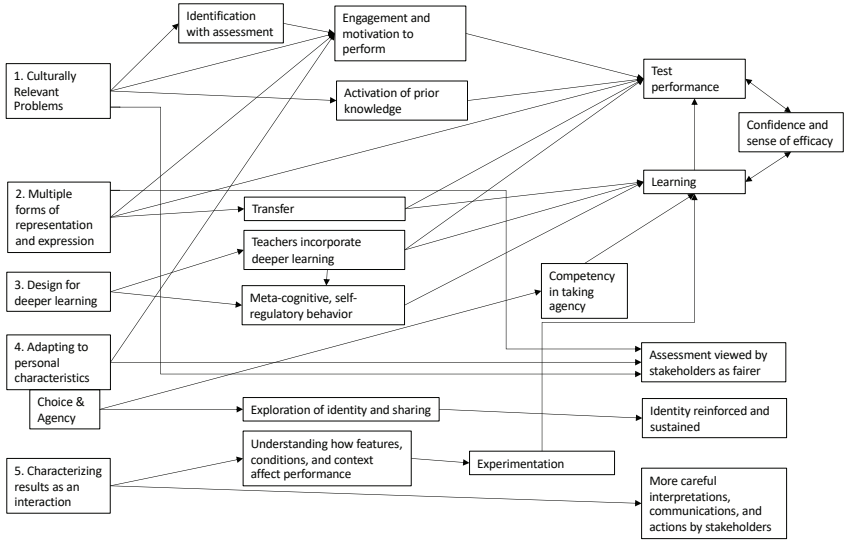
To date, the evidence on adaptation to personalize learning and assessments is scattered, based on partial information, small sample trials, and particularized settings. How might one determine whether an assessment that was personalized for the advancement of equity and learning was effective? Bennett (2023) proposed a theory for socioculturally responsive assessment that offers one approach to addressing this question. The theory is built from five assessment design principles drawn from multiple literatures, including those on the teaching of traditionally underserved students, the assessment of such students, and the learning sciences. The principles are as follows:

1. Present problem situations that connect to, and value, examinee experience, culture, and identity because individuals are more likely to show what they know and can do in familiar versus unfamiliar situations;
2. Allow for multiple forms of representation and expression in problem stimuli and in responses, providing more equitable access to question content and to paths for response;
3. Promote instruction for deeper learning through assessment design to increase the likelihood of exposing students in lower-performing schools to such learning;
4. Adapt the assessment to student characteristics, including by offering choice of problems and agency in contributing to problem definition; and
5. Represent assessment results as an interaction among what the examinee brings to the assessment, the types of tasks engaged, and the conditions and context of that engagement, thereby giving a more nuanced depiction of performance.

As should be clear, except for promoting deeper learning through assessment design, each of these principles directly implies personalization: presenting problems that connect to examinee experience, culture, and identity; allowing forms of representation and expression suited to the examinee; adapting the assessment to student characteristics; and framing results to reflect the unique intersection of factors brought by the examine and by the assessment. Designing the assessment to promote deeper learning is aimed at encouraging adaptations that maintain rigor.

Figure 1 shows how these design principles work together theoretically to produce their intended effects. In the Figure, the theory is represented as a network of empirically testable propositions emanating from the design principles. On the left are the five principles. To their right are intermediate and ultimate outcomes (i.e., the postulated effects of designing and implementing an assessment according to those principles). In terms of the theory, a personalized assessment is effective to the degree that a preponderance of the posited propositions are supported by the empirical evidence. Next, the principles are listed, in shortened form, each followed by a discussion of the propositions, along with the types of studies that might be used to evaluate them. Because it is not particular to personalized assessment, principle #3 (promote instruction for deeper learning …) is not discussed.

Figure 1.
An initial theory of socioculturally responsive assessment



Note. Adapted from "Toward a theory of socioculturally responsive assessment" (p. 97), by R. E. Bennett, 2023, *Educational Assessment, 28*(2). Copyright [2023] by Educational Testing Service. Adapted by permission.

Principle 1: Present problem situations that connect to examinee experience, culture, and identity

The propositions emanating from this principle posit that a learner's identification with an assessment will increase when the assessment incorporates problems connected to their cultural identity, background, and lived experiences. Increased identification should in turn promote higher levels of engagement and motivation to perform. In addition, the incorporation of culturally relevant problems should help to activate prior knowledge, causing students to perform better than on problems that do not make such personal linkages. Higher levels of performance should raise confidence and bolster sense of efficacy, thereby enhancing learning and future test performance. Lastly, stakeholders should be led to feel that the assessment is fairer.

This proposition set might be tested through studies that experimentally manipulated the cultural relevance of problems for individuals, examining the degree to which aligned problems resulted in the described effects: for examinees, increased identification, engagement and motivation to perform, activation of prior knowledge, and performance; for stakeholders, improved perceptions of fairness.

Principle 2: Allow for multiple forms of representation and expression in problem stimuli and in responses

Students should be better able to show what they know and can do when they are offered multiple forms of representation (e.g., essay, bulleted list, drag-and-drop graphical form) and allowed alternate modes of expression (e.g., spoken, written, sign). Such affordances should cause increased student performance and perceptions among stakeholders that testing is fair. When assessment incorporates problems that encourage making deep-structure connections among representational forms and expressive modes, subsequent transfer of learning should be enhanced and test performance improved.

The propositions associated with this principle might be tested through studies that experimentally manipulated the allowance of single versus multiple forms of representation and single versus multiple forms of expression. For example, in the case of a science content assessment, students might be randomly assigned to a condition that permitted choice of representational form (e.g., brief essay, bulleted list, drag-and-drop graphical form) versus one that required responding only via a brief essay. Along with the main effects on engagement, motivation,

and performance, in a sample large enough to disaggregate demographically, the group-by-condition interaction could also be evaluated. Such studies should evaluate whether student-made choices actually benefited performance or are made on more superficial considerations, for instance, to shorten testing time (See Powers & Bennett, 1999, for a review of research on choice).

Principle 4: Adapt the assessment to student characteristics

The propositions linked to this principle concern the postulated effects of adjusting to personal characteristics. Such adaptation should better align the assessment with student interests, cultural identity, background, and prior knowledge than does a traditional test, thereby causing stakeholders to feel the assessment is fairer. Higher levels of motivation and engagement with the test should also result, thereby enhancing performance. With teacher guidance in making similar choices over the school year, permitting choice should aid the development of student competency in taking agency. That competency should, in turn, impact learning positively. Moreover, when allowing agency motivates examinees to examine their identity and share those explorations, identity should be reinforced and sustained. Especially for minoritized students, the positive effects of adaptation should increase with the degree to which the assessment meaningfully takes personal characteristics into account.

Among other studies, members of this proposition set might be evaluated by asking stake holders to rate the fairness of hypothetical examples of adaptation to particular student characteristics, as well as non-adaptive examples. In addition, asking students to rate engagement and motivation to perform, after taking personalized versus non-personalized tasks designed to measure the same content, would be informative, as would evaluating any performance differential and its direction. Questions of comparability might also be evaluated by correlating scores from both traditional and personalized assessments with measures of criterion performance, and by structural equation modeling approaches.

Principle 5: Represent assessment results as an interaction among the examinee, the tasks, and the conditions and context

The propositions implied by this principle postulate that examinees, teachers, the public, and policy makers will more carefully interpret, communicate about, and act on assessment results if those results are represented to them as an interaction among examinee characteristics, the features of tasks, and the conditions and context of the assessment. Because the notion of fairness has been embedded in uniformity, strong evidence of the utility of adaptations should be required. In this regard, more careful interpretation implies viewing results as closely tied to the task types, conditions, and contexts that characterized the assessment. When results are thought of in this way, students should be more likely to make connections between how task features, conditions, and contexts affect their performance. Additionally, teachers and students should be led to explore adjustments to these factors to see how they might impact learning and improve test performance. Studies that provide a treatment to teachers and students designed to sensitize them to this interactional perspective, followed by cognitive interviews and classroom observations, might help uncover whether the postulated effects emerge.

Conclusion

Our thesis has been simple, that assessments (and learning) should be designed and implemented to provide flexibility and adaptation to individual differences. This orientation makes demands on many groups. Some of the burden falls on test developers who must deal with the complexities of implementation. Teachers' engagement will also be essential if students are to be given differential types of learning and assessments, in particular to equally value those types and communicate that perspective to students. Students are also encouraged through metacognitive and self-regulated thinking to become more thoughtful about the choices they make in assessment contexts (Bembenutty, this volume). Policy makers and parents need preparation and examples so they can accept assessments that differ among students.

A tremendous burden falls upon researchers and the schools and districts which must help in generating the evidence. Evidence needs to be generated to explore the utility of the model systematically and to provide guidance for its revision, as needed. Teacher and student participants need to be available, across individual

and group variations, subject matter domains, skills, and age ranges. Perhaps the evaluation strategy should start with an age range, such as that found in middle-school, and a particular subject matter. If positive evidence is found, a second step might be to evaluate the effects in other domains and other age levels. Whatever the particular research strategy, it is clear that research into the major premises of this work will ultimately need to be supported by government, commercial test developers, purveyors of instructional systems, and teacher development organizations. The problem of endemic differential performance must be finally addressed, and it will take more than a village to do so.

References

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm.* Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

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 ed.). Longman.

Atkinson, R. C. (1972). Ingredients for a theory of instruction. *American Psychologist, 27*(10), 921–931.

Baker, E. L., Sawaki, Y., & Stoker, G. (2002, April 1–5). *The LEARNOME: A descriptive model to map and predict functional relationships among tests and educational system components* [Paper presentation]. Annual meeting of the American Educational Research Association, New Orleans, LA, United States.

Bandura, A. (1977). *Social learning theory.* Prentice-Hall.

Bellamy-Walker, T. (2022, January 18). *NJ mandates teaching Asian American and Pacific Islander history in schools*. NBC News. https://www.nbcnews.com/news/asian-america/nj-mandates-teaching-asian-american-pacific-islander-history-schools-rcna12637

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. (2025). A descriptive review of culturally responsive, socioculturally responsive, and related assessment conceptions. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 11–28). Routledge. https://doi.org/10.4324/9781003435101

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* David McKay.

Brummelman, E., & Sedikides, C. (2023). Unequal selves in the classroom: Nature, origins, and consequences of socioeconomic disparities in children's self-views. *Developmental Psychology, 59*(11), 1962–1987.

Buzick, H. M., Casabianca, J., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice.* https://doi.org/10.1111/emip.12551

Cavanaugh, S. (2023, February 27). State restrictions like Florida's on curriculum are having an effect nationwide. *EdWeek Market Brief.* https://marketbrief.edweek.org/regulation-policy/state-restrictions-like-floridas-on-curriculum-are-having-an-effect-nationwide/2023/02

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum, 1*(8) 139–167. http://chicagounbound. uchicago.edu/uclf/vol1989/iss1/8

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

De Bruyckere, P., Kirschner, P. A., & Hulshof, C. D. (2015). *Urban myths about learning and education.* Academic Press.

Durán, R. (2022, November 1–2). *Cultural context and affirmative pedagogies* (Facilitator). Extending His Reach: Promoting the Educational Legacy of Edmund W. Gordon, hosted by the CRESST, UCLA School of Education & Information Studies, and Gevirtz School of Education, UC Santa Barbara, Los Angeles, CA, United States.

Gagné, R. M., & Medsker, K. L. (1996). *The conditions of learning: Training applications*. Harcourt Brace.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116.

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice* (3rd ed.). Teachers College Press.

Gordon, E. W., Boykin, A. W., Armour-Thomas, E., & McCallister, C. (2019). *Human variance and assessment for learning*. Third World Press Foundation.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124.

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

Hattie, J. (2023). *Visible learning: The sequel. A synthesis of over 2,100 meta-analyses relating to achievement.* Routledge.

Hobbs, F., & Stoops, N. (2002). *Demographic trends in the 20th century: Census 2000 special reports* (CENSR-4). US Census Bureau. https://eric.ed.gov/?id=ED477270

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education, 67*(3), 187–196. https://doi.org/10.2307/2668188

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. The WAC Clearinghouse; Parlor Press. https://doi.org/10.37514/PER-B.2015.0698

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2).

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491.

Ladson-Billings, G. (2021). *Culturally relevant pedagogy: Asking a different question.* Teachers College Press.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education, 67*(3), 268–279.

Lee, C. D., Meltzoff, A. N., & Kuhl, P. K. (2020). The braid of human learning and development: Neurophysiological processes and participation in cultural practices. In N. S. Nasir, C. D. Lee, R. Pea, & M. McKinney de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 24–43). Routledge.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). *A call to action: Confronting inequity in assessment.* Lyons Assessment Consulting. https://www.lyonsassessmentconsulting.com/resources.html

Mayer, R. E. (2014). Incorporating motivation into multimedia learning, *Learning and Instruction, 29*, 171–173.

Mays, M. (2021, October 14). California students will have to take ethnic studies to get a diploma. *Politico*. https://www.politico.com/news/2021/10/14/california-students-ethnic-studies-diploma-515972

Merrill, M. D. (2009). First principles of instruction. In C. M. Reigeluth & A. Carr (Eds.), *Instructional design theories and models: Building a common knowledge base* (Vol. III). Routledge.

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

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into Practice, 13*(2), 132–141.

National Center for Education Statistics (NCES). (2020). *Digest of education statistics: 2020 tables and figures.* US Department of Education. https://nces.ed.gov/programs/digest/d20/tables/dt20_203.70.asp

Oakes, J. (2005). *Keeping track: How schools structure inequality.* Yale University Press.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 425–441). Pergamon.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review, 31,* 339–359.

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education, 12(3)*, 257–279. doi:10.1207/S15324818AME1203_3

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology, 29*, 417–447.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90. https://doi.org/10.1111/emip.12429

Rana, R., & Mahmood, N. (2010). The relationship between test anxiety and academic achievement. *Bulletin of Education and Research, 32*(2), 63–74.

Rogoff, B. (2003). *The cultural nature of human development.* Oxford University Press.

Rogoff, B., & Angelillo, C. (2002). Investigating the coordinated functioning of multifaceted cultural practices in human development. *Human Development, 45*, 211–225.

Shute, V. J., D'Mello, S. K., Baker, R., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457–490.

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The Matrix of evidence for validity argumentation. *Frontiers in Education, 4*(43). https://doi.org/10.3389/feduc.2019.00043

US Census Bureau. (2021). *Quick facts: United States*. https://www.census.gov/quickfacts/fact/table/US/PST045219

Vertovec, S. (2023). *Superdiversity: Migration and social complexity.* Routledge.

Ward, W. C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 271–282.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education* (Research Monograph Number 6). CCSSO.

Weiss, D. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1–27.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge University Press.

A Theory-Informed and Student-Centered Framework for Comprehensive Educational Assessment

Norris M. Haynes, Mary K. Boudreaux, and Edmund W. Gordon

Abstract

This chapter presents a comprehensive theory-informed and student-centered framework for educational assessment. This framework is grounded in three theoretical perspectives: constructivism, sociocultural theory, and implicit theory. Constructivism emphasizes learners' active roles in knowledge construction and the use of assessments to understand their thinking processes. Sociocultural theory emphasizes how social and cultural elements shape the learning process, underscoring the importance of comprehending the educational environment and its effects on student growth. Implicit theory addresses assumptions about individual abilities, distinguishing between the entity (fixed) and incremental (malleable) perspectives. The framework also considers essential student characteristics such as developmental stages, demographic backgrounds, motivation, and cultural contexts, which profoundly influence student engagement and performance. The framework accounts for the curriculum and instructional context, including the hidden curriculum that shapes the school and classroom environments. This chapter explores various assessment approaches, including formative, summative, diagnostic, ipsative, norm-referenced, criterion-referenced, curriculum-based, self-assessment, and holistic methods. These diverse strategies provide a comprehensive understanding of student learning and support individualized growth. The integration of theoretical principles, student characteristics, and contextual

factors into the assessment design fosters a supportive, equitable, and growth-oriented learning environment. Ultimately, this theory-informed and student-centered assessment framework promotes meaningful learning experiences, academic excellence, and the holistic development of all students.

Introduction

A vast body of literature on student assessment has been produced by scholars over the years, including numerous manuscripts authored by Professor Edmund Gordon and his colleagues, some of which are highlighted in the Gordon Commission's report on assessment (Armour-Thomas & Gordon, 2012; Gordon et al., 2012, 2013, 2016; Bereiter & Scardamalia, 2012). Given the existing assessment literature, we view the invitation to contribute to a chapter on assessment as an opportunity to present innovative ideas and a unique perspective, along with a theory-informed and student-centered assessment framework that has practical applications. This chapter focuses on a broad audience of educators, including teachers, educational leaders, and policymakers. The goal of this chapter is to provide a comprehensive assessment framework that integrates various factors that support student learning. These factors must be woven into a cohesive assessment system to ensure that evaluations of student performance are grounded in a thorough understanding of the multifaceted and complex educational environment in which students operate. Thus, assessments evolve from mere tools for measuring knowledge to instruments that enhance and enrich holistic educational experiences.

Our assessment framework consists of four pillars. First, our theory-informed and student-centered assessment framework is grounded in three approaches. Constructivism emphasizes the active participation and involvement of students in their learning journey. This approach also uses evaluations to gain insight into how learners develop their understanding and build knowledge structures. Sociocultural theory highlights the impact of social and cultural factors on learning and includes an understanding of the sociocultural learning context and its effects on students' development and learning. Implicit theory addresses assumptions regarding individual abilities. Implicit theory encompasses two main perspectives: entity and incremental perspectives.

Second, our framework requires attention to student characteristics empirically related to learning and achievement. These characteristics include developmental and demographic background variables and students' motivational factors.

Third, our framework considers the curriculum and instructional context in which teaching and learning occur, including the hidden curriculum, which is part of the culture and climate of schools and classrooms. The hidden curriculum includes implicit lessons and values conveyed within the educational environment that influence students' socialization and development (Jackson, 1968).

Finally, we consider various assessment approaches and methods for evaluating how students demonstrate learning and intellectual growth. We believe student-centered assessments are most useful, effective, and valid when designed within a comprehensive framework informed by these three theoretical frameworks: constructivism, sociocultural theory, and implicit theory. In doing so, we align the goals of this chapter with the following Principles for Assessment in the Service of Learning: (3) assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition; (4) assessments model the structure of expectations and desired learning over time; (5) feedback, adaptation, and other relevant instruction should be linked to assessment experiences; (6) assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences; and (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.

Figure 1 at the end of this chapter captures how our three primary theoretical frameworks—constructivism, sociocultural theory, and implicit theory—form the foundation for assessment approaches: formative, summative, diagnostic, ipsative, norm-referenced, criterion-referenced, curriculum-based, self-assessment, and holistic. The figure also incorporates the essential elements of student characteristics, curriculum, and instruction context factors. By visually connecting theoretical frameworks, assessment approaches and essential elements, the figure reinforces the comprehensive and integrated nature of student assessment discussed throughout the chapter.

Theoretical Frameworks

To develop a robust understanding of student assessment, it is essential to anchor discussions within well-established theoretical frameworks. These theories provide the foundation for interpreting and designing meaningful and effective assessment practices. In this section, we examine our assessment framework through the lens of constructivism, sociocultural theory, and implicit theory. These perspectives address the diverse needs of students and the complex contexts in which learning occurs and offer dynamic and equitable assessment approaches. Each framework provides unique insights into how assessments can promote student growth, equitable practices, and deeper engagement, thereby forming an integrated foundation for meaningful evaluation and learning.

Constructivism

Constructivist theories emphasize the importance of active student-centered learning. Vygotsky's (1978, 1986) work highlights how learners construct meaning through developmental and phenomenological experiences, positioning their involvement at the center of both learning and assessment. Authentic assessment systems within this framework integrate subjective and objective measures to evaluate how effectively students apply their knowledge to complex, real-world problems.

Piaget's (1952, 1970) cognitive development theory further underscores that learners actively construct knowledge through meaningful interactions with their environment. Piagetian assessments align with students' developmental stages, prioritizing their cognitive and problem-solving processes (Snow, 1977). Similarly, Dewey (1938) champions experiential learning and advocates assessments that measure how meaningful and relevant educational experiences are for individual learners. Together, these perspectives emphasize tailoring assessments of students' developmental, experiential, and contextual needs of students to ensure a more authentic evaluation of their abilities.

Assessment Implications of Constructivism Theory

Constructivism principles form the foundation for creating assessment methods that extend beyond traditional standardized tests. These techniques aimed to evaluate students' capacity to utilize their knowledge in intricate, practical scenarios. In this context, authentic assessment systems integrate subjective

and objective evaluations, prioritizing critical thinking, innovation, and problem-solving abilities over rote memorization. These assessments also underscore the importance of actively constructing knowledge and promoting deeper comprehension and skill development.

The constructivist approach provides a more comprehensive and genuine perspective on learning by customizing assessments to align with students' developmental requirements, backgrounds, and environments. These evaluations not only gauge learning outcomes but also cultivate higher-order thinking, teamwork, and self-reflection, equipping learners with transferable skills essential for navigating complex challenges in both academic and real-world settings. The principles and practices of constructivism highlight the adaptability of constructivist assessments and their capacity to generate meaningful student-centered educational experiences.

Sociocultural Theory

Diversity in today's classrooms profoundly shapes children's lives, influencing their identities, learning processes, and methods of expressing knowledge. Vygotsky's sociocultural theory (1978, 1987, 1986) calls for attention to the essential role of social and cultural environments in the learning process, stressing how cognitive development is affected by interpersonal connections and social interactions. Sociocultural assessments reflect these contexts by incorporating collaboration skills, communication, and interaction and transforming evaluations into tools that foster social and cultural awareness.

Sociocultural assessment is vital for English learners. They integrate language learning, literacy, and fluency to ensure that evaluations address the diverse needs of students. Sireci's (2020) concept of "understandardization" challenges the rigidity of traditional testing, advocating for greater flexibility in accommodating diverse learners. This approach ensures that standardized testing accounts for unique needs while maintaining validity and creating adaptive environments that reflect the diversity of the student populations.

Assessment Implications of Sociocultural Theory

Sociocultural assessment methods consider learners' experiential background and social context. Assessments include culturally sensitive criteria, such as evaluating written assignments, observing social interactions, and analyzing language. These practices ensure that assessments are academically rigorous while respecting and reflecting students' diverse cultural and social backgrounds. By embracing culturally responsive approaches, educators can create assessments that address all students' holistic learning experiences of all students.

Implicit Theory

Implicit theories focus on individuals' beliefs about their abilities and their potential for growth. Integrating these theories into teaching and assessment fosters a positive and motivational environment. As a result, learners develop a heightened sense of empowerment and influence over their educational journey, leading to improved academic performance, increased confidence in their abilities, and enhanced internal drive to learn. Dweck's (2006) mindset theory provides the foundation for understanding these dynamics. In this section, we focus on two key implicit theories: (a) mindset theory and (b) attribution theory.

Mindset theory. Dweck's (2006) Mindset Theory distinguishes between fixed and growth mindsets. A fixed mindset aligns with the entity perspective of implicit theories, viewing intellectual capacities as static. Conversely, a growth mindset reflects an incremental perspective, emphasizing that abilities develop through effort, practice, and learning. Dweck's research demonstrates that students with growth mindsets display persistence, perseverance, and sustained effort, leading to more successful outcomes than those with fixed mindsets who often experience challenges as insurmountable.

Assessment Implications of Mindset Theory

Evaluations that align with a growth-oriented mindset prioritize ongoing improvement techniques and formative feedback. These types of assessments motivate students to perceive errors as opportunities to learn, thereby promoting their involvement, determination, and confidence in their capacity for development and growth (Cauley & McMillan, 2010). By reinforcing effort and improvement strategies, educators shift the focus from static to dynamic, process-driven learning, cultivating resilience and a passion for lifelong learning.

Attribution theory. Weiner's (1974, 1985) Attribution Theory examines how individuals assign causes to success and failure. This theory highlights three key dimensions: locus of control (internal versus external factors), stability (stable versus unstable causes), and controllability (controllable versus uncontrollable factors). These dimensions influence students' perceptions of their abilities and their potential for success.

Assessment Implications of Attribution Theory

Attributional strategies, such as retraining and reframing failure, encourage students to attribute successes to effort and effective strategies and foster control over outcomes. Reflective assessments further enable students to analyze their learning processes and connect their efforts to their achievements. Thomas Edison's concept of "intelligent failure" promotes resilience by framing setbacks as valuable learning opportunities (Edmondson, 2023). Together, these strategies cultivate self-efficacy and proactive learning approaches.

Application of Theories to Assessment

The theoretical frameworks discussed thus far have provided valuable insight into how students learn and develop. This section explores concrete examples of how these theories can be applied to assessment practices in an educational setting. By examining specific applications, we can understand how constructivism, sociocultural theory, mindset theory, and attribution theory can inform and enhance the assessment approaches. These examples demonstrate how theory-informed assessments can provide more comprehensive, equitable, and growth-oriented evaluations of students' learning and progress. Through these practical illustrations, educators can develop ideas for implementing theory-informed assessment strategies in classrooms and schools.

Constructivism

Constructivist principles emphasize active, student-centered learning and provide a foundation for designing meaningful and practical assessments. These assessments exceed traditional standardized tests by focusing on critical thinking, problem-solving, and hands-on learning. The objective is to assess students' comprehensive knowledge and their capacity to implement it in realistic and practical scenarios.

Example: Inquiry-based science projects. In an elementary school science class, students may be tasked with designing and conducting experiments to investigate factors affecting plant growth. Groups can explore variables such as light conditions, soil types, and watering schedules, hypothesizing their effects on seed germination and plant development. Throughout the project, students document their observations, collect data, and analyze results, culminating in a report or presentation to share their findings. These assessments extend beyond testing memorized facts by evaluating students' abilities to apply scientific methods, interpret data, and reason through unexpected outcomes. Teachers evaluate not only the results but also the entire journey, including students' teamwork abilities, their approach to problem-solving, and their resilience when confronted with obstacles. By emphasizing these elements, this approach reflects constructivist principles, fostering active knowledge construction, deeper understanding, and the development of transferable skills essential for real-world scientific inquiry.

Implications for constructivist assessment practices. Constructivist assessments such as portfolios, project-based tasks, and collaborative activities demonstrate how constructivist principles translate into effective educational practices. For example, portfolios provide a longitudinal perspective on students' progress, highlighting reflection and the application of learning across multiple contexts. Unlike inquiry-based projects that emphasize collaboration and experimentation, portfolios allow students to document individual growth and showcase their understanding of various subjects. These complementary approaches illustrate the flexibility of constructivist assessments to address diverse learning goals.

By valuing both the process (e.g., collaboration, problem-solving, adaptability) and the product (e.g., final reports, portfolios), constructivist assessments foster higher-order thinking, metacognitive reflection, and adaptability. These methods equip students with transferable skills such as critical analysis, creativity, and resilience, preparing them for complex, real-world challenges. Together, these assessment strategies highlight the versatility of constructivist approaches in promoting meaningful, student-centered learning experiences that go beyond traditional evaluation methods.

Sociocultural Theory

Sociocultural theory posits that the learning process is fundamentally influenced by culture and social interactions among individuals (Vygotsky, 1978). Building on this foundation, assessments that incorporate collaborative activities and peer evaluations allow educators to observe how students engage with one another and apply cultural tools to learning environments. For instance, a group project might require students to collaboratively solve real-world problems by integrating diverse perspectives and cultural understanding (Hambleton & Zenisky, 2011). Teachers can assess not only the final product but also the dynamics of the group's interaction, such as communication, negotiation, and conflict resolution. This approach aligns with Vygotsky's focus on learning through social collaboration and provides deeper insights into students' abilities to navigate and thrive in socially constructed learning environments.

Example: In a classroom with diverse learners, a teacher can design assessments that incorporate group activities in which students collaborate to solve problems. These assessments not only evaluate students' knowledge but also their ability to communicate and work effectively with peers (Reeves, 2017), reflecting Vygotsky's emphasis on social interaction as a key to learning. Such collaborative assessments involve complex real-world scenarios that require students to apply their knowledge across multiple disciplines to foster critical thinking and problem-solving skills. For instance, students could be tasked with developing a sustainable urban planning project that requires them to consider the environmental, economic, and social factors. Additionally, these group assessments can be structured to include roles that cater to different learning styles and strengths, ensuring that every student meaningfully contributes. For instance, one student might excel in research and data analysis, whereas another might be skilled in visual presentations or public speaking. Through the allocation of distinct roles within the group, the teacher cultivates a learning atmosphere that embraces diverse skills and insights, thus promoting greater inclusivity.

Implications. This example highlights the transformative potential of assessments informed by sociocultural theory, particularly in leveraging social interactions and cultural diversity as central to the learning process. By integrating culturally responsive assessments, such as oral presentations, collaborative projects, and visual representations, educators can address the diverse learning needs of students, especially English Language Learners (ELLs) (Solano-Flores, 2013).

The project centered on sustainable urban design can challenge students in multicultural teams to tackle real-world issues. This approach promotes analytical thinking, creative problem-solving, and the application of theoretical knowledge in authentic situations. These assessments foster inclusivity by evaluating students' unique cultural perspectives and diverse learning approaches (Bryant et al., 2019).

Incorporating flexible assessment practices, as advocated by Sireci's (2020) concept of "understandardization," further enhances equity by accommodating the diverse needs of students while maintaining validity. For instance, allowing multilingual students to present findings in their preferred language (translanguaging; see García & Wei, 2014) or format ensures that assessments capture their full capabilities rather than penalizing them for language barriers.

By organizing group assessments to reflect students' individual strengths and roles, educators can ensure that a variety of talents and perspectives are acknowledged, fostering engagement and a sense of belonging. This approach embodies sociocultural principles by positioning learning as a collective, contextually grounded process (Diaz & Berk, 1992). As a result, students are better prepared to navigate complex, multicultural environments and develop the collaboration and adaptability needed for success in an increasingly interconnected world.

Mindset theory and growth-oriented assessment. Dweck's (2006) mindset theory emphasizes the importance of fostering a growth-oriented mindset among students. This approach encourages learners to view their abilities as adaptable and capable of improvement through dedicated effort and persistence. This belief fosters better performance, resilience, and a willingness to embrace challenges. While the theory's foundational principles have transformed educational psychology, its most significant impact lies in classroom practices that emphasize growth-oriented assessments. These methods encourage students to concentrate on improvement and strategy rather than fixed outcomes, reinforcing the idea that learning is an ongoing process.

Example: Incremental quizzes in high school math. In high school math classes, incremental quizzes provide students with regular opportunities to receive feedback regarding their progress. These quizzes serve as checkpoints, helping students identify their strengths and weaknesses in specific topics while alleviating the

stress linked to high-stakes final exams. Teachers can utilize the quiz results to clarify misconceptions, reinforce difficult concepts, and tailor their instruction to meet students' needs.

Unlike traditional assessments, incremental quizzes allow students to revisit topics and improve them over time, cultivating a mindset in which mistakes are viewed as valuable learning opportunities. For instance, students may correct and resubmit their quizzes as part of the learning process and receive additional feedback to guide their efforts. This iterative approach not only builds confidence but also strengthens critical thinking and problem-solving skills.

Implications for growth-oriented assessment. Growth-oriented assessments, such as incremental quizzes, formative assessments with iterative feedback, and self-reflection activities foster a classroom culture aligned with Dweck's (2006) principles. These methods emphasize effort, strategy, and persistence, encouraging students to analyze their mistakes, refine their approaches, and celebrate incremental progress. By valuing growth over static performance, such assessments promote inclusivity and ensure that every student succeeds.

Furthermore, this approach prepares students for real-world challenges where resilience and the ability to adapt and learn from mistakes are critical. For example, reflection activities that require students to analyze their problem-solving processes can enhance metacognitive skills, equipping them to tackle complex, unpredictable tasks beyond the classroom. By embedding these practices into assessment design, educators can create supportive environments that nurture lifelong commitments to learning and personal growth.

Attribution Theory

Weiner's Attribution Theory (1974; 1992) explores how individuals perceive the causes of their successes and failures, focusing on three dimensions: locus of control (internal versus external factors), stability (stable versus unstable causes), and controllability (controllable versus uncontrollable factors). This theory underscores the motivational benefits of attributing outcomes to internal and controllable factors such as effort, rather than external circumstances or inherent ability. When students believe that their success is tied to their actions, they are more inclined to overcome obstacles and assume ownership of their learning (Weiner, 2005).

Assessments based on attribution theory encourage students to focus on their capacity for growth and improvement. This is achieved through methods that emphasize the effort and strategic application of feedback, reinforcing students' sense of control and personal agency during their learning journey.

Example: Revising cognitive development is essential in a psychology course. In a college psychology course, assessments designed using Attribution Theory include incremental feedback and structured reflection opportunities. For example, students are assigned an essay on cognitive development theories and asked to submit a first draft for detailed feedback. The professor highlighted the strengths and areas of improvement in content, argument structure, and citation practices, providing actionable guidance for revision.

Students are then required to reflect on feedback in the journal, document their revision process, and describe how they plan to address each critique. To further support reflection, optional peer review sessions are offered, allowing students to gain diverse perspectives and refine their work collaboratively. Upon resubmission, the students included a summary of the changes they made and their rationale, which the professor evaluated alongside the revised essay. This iterative process highlights the importance of effort, strategy, and adaptability, reinforcing the belief that academic success is largely within control.

Implications for assessment design. Attribution theory offers a unique lens for designing assessments that foster personal growth, resilience, and self-regulation. Unlike other approaches, it specifically emphasizes controllability, showing students how their efforts directly affect their outcomes. For example, actionable feedback and structured opportunities for revision empower students to connect diligence with improved performance. By engaging in this process, students develop a deeper understanding of their role in shaping their learning outcomes, promoting an internal locus of control, and enhancing their academic self-confidence.

Moreover, assessments informed by attribution theory prepare students for real-world challenges by teaching them to analyze feedback, strategize effectively, and learn from their mistakes. This approach aligns with the broader educational goals of cultivating critical thinking, problem-solving, and adaptability, equipping students with skills essential for lifelong learning and success (Cronbach & Snow, 1977).

Summary of Theoretical Frameworks into Assessment Practices

Integrating theoretical frameworks into assessment practices provides a comprehensive approach to evaluating and supporting student learning. Constructivism shapes assessments that engage students in actively constructing knowledge and emphasizing practical activities, critical thinking, and teamwork. By addressing students' developmental and experiential needs, these methods promote analytical thinking and a more comprehensive understanding of ideas.

The sociocultural perspective introduces an essential aspect by acknowledging how cultural and social environments influence the learning processes. Assessments grounded in this framework prioritize inclusivity and responsiveness, accommodating diverse perspectives and encouraging culturally meaningful evaluation methods that support all learners.

Mindset and attribution theories further enhance assessment practices by focusing on student motivation and self-perception. The principles of a growth mindset inform the creation of assessments that emphasize effort and progress, motivating students to perceive challenges as avenues for personal growth and skill enhancement. Attribution theory enhances this approach by encouraging students to associate their achievements with factors within their control such as diligence and methodological choices, thereby cultivating a sense of empowerment and adaptability in their educational pursuits.

Together, these frameworks create a holistic, student-centered approach to assessment that promotes equity, supports diverse learning needs, and prepares students for long-term academic and personal success. By integrating these perspectives, educators can design assessments that not only measure performance but also inspire growth, adaptability, and lifelong commitment to learning.

Student Characteristics

Building on the foundation of theoretical frameworks, understanding the diverse characteristics that students bring to a learning environment is critical for creating effective and equitable assessment practices. Student characteristics, including developmental stages, demographic backgrounds, prior knowledge, motivation, and cultural contexts, profoundly influence how students engage with assessments and interpret their outcomes. Recognizing and addressing these individual differences

allows educators to design assessments that support each student's unique learning journey while ensuring inclusivity and responsiveness.

Significance of Student Characteristics in Assessment

Research has highlighted the need for performance assessments that account for the breadth of student abilities. Shavelson, Baxter, and Pine (1992) emphasized the importance of considering factors such as students' backgrounds, experiences, and learning styles in assessment design. Tailoring assessments of these diverse characteristics ensures that they accurately reflect students' learning and engagement (Bryan et al., 2019). Sireci's (2020) concept of "understandardization" further advocates for assessments that accommodate varied student attributes, such as cultural background, learning styles, and subjective experiences, thereby enhancing their validity and equity.

Key Influences on Learning and Assessment

Socioeconomic Status (SES). Sirin's (2005) meta-analysis demonstrated a significant relationship between SES and academic achievement, with students from higher SES backgrounds often outperforming their peers from lower SES backgrounds. This underscores the importance of considering socioeconomic factors in designing equitable assessment strategies.

Learning style and instructional approach. While Coffield et al. (2004) highlighted the benefits of tailoring instruction to diverse learning styles, they also cautioned against overreliance on this approach. Instead, they advocated a holistic understanding of student diversity by incorporating multiple dimensions of individual differences into educational practices.

Motivation and engagement. Motivation is a central driver of academic success. Pintrich & De Groot (1990) and Deci et al. (1991) emphasized the critical role of intrinsic motivation and self-regulation in achieving academic goals. Self-determination theory suggests that students who feel autonomous and competent are more likely to engage deeply in their learning. Similarly, Fredricks et al. (2004) demonstrated that student engagement significantly impacts academic outcomes, highlighting teachers' role in fostering a motivating and engaging learning environment.

Cultural and linguistic diversity. Valdés (1996) explored how cultural and linguistic factors shape educational experiences, particularly in bilingual and multilingual contexts. Effective assessments must account for these factors to ensure inclusivity and relevance for students from diverse cultural backgrounds.

Cognitive load and capacity. Sweller's (1988) Cognitive Load Theory emphasizes the importance of balancing task complexity with students' cognitive capacity. Overloading students can hinder their ability to process information effectively, thus underscoring the need for assessments that match their cognitive development levels.

Self-efficacy and goal setting. According to Zimmerman et al. (1992), academic motivation is significantly shaped by an individual's self-efficacy beliefs and their ability to set goals. Students with high self-efficacy and clear self-determined goals are more likely to persevere through challenges, achieve higher academic standards, and experience success.

Implications for Practice

Educators must consider the interplay between these diverse student characteristics when designing effective teaching and assessment strategies. Assessments that align with students' socioeconomic contexts, cultural and linguistic backgrounds, cognitive abilities, and motivation levels can foster more inclusive and equitable educational environments. As an illustration, evaluative techniques such as portfolio collections or inquiry-driven assignments offer students the opportunity to exhibit their academic progress in a manner that accentuates their distinctive capabilities and lived experiences.

By incorporating these factors into assessment design, educators can create practices to evaluate academic achievement and promote critical thinking, resilience, and lifelong learning. This multifaceted approach ensures that education is responsive to every student's needs and potential, enhances overall academic success, and fosters an inclusive and supportive learning environment.

Curriculum and Instructional Context Factors

This section examines how educational assessment is an integral component of the interconnected triad of Curriculum, Instruction, and Assessment (CIA), all of which are fundamental to the teaching and learning process. The curriculum defines the content that students are expected to learn, instruction encompasses the methods used to deliver that content, and assessment evaluates the effectiveness of both in fostering student learning and development. Effective assessments provide actionable insights that enhance instructional strategies, refine curricular goals, and improve student outcomes (Brown & Miller, 2021). Without well-designed and implemented assessment processes, the purpose and impact of the curriculum and instruction are significantly diminished.

The value of teaching and learning experiences is greatly enhanced when assessments are designed to meet student needs. Ainscow et al. (2006) emphasized the importance of creating inclusive teaching and learning environments, a principle that has profound implications for assessment practices. Assessment should move beyond its traditional summative role of measuring outcomes and instead function as a dynamic tool for monitoring, addressing, and enhancing student learning and growth. By adopting this approach, assessment becomes a mechanism for understanding not only the results but also the ongoing developmental journey of each learner. This shift in focus encourages a holistic view of education, prioritizing continuous improvement over static evaluations.

The Impact of Instructional Context Factors

Instructional context factors, although frequently discussed, have rarely been assessed comprehensively. These factors include both visible and less visible elements of the learning environment, such as classroom culture, teacher-student interactions, and sociocultural practices. Drawing from Hall's (1976) cultural iceberg theory, we recognize that school and classroom contexts consist of both surface- and deep-level structures. The surface structure, analogous to the visible tip of an iceberg, includes explicit curricula, instructional strategies, classroom policies, and observable behaviors. These are overt and immediately recognizable components of the educational environment.

By contrast, the deep structure represents the hidden curriculum—the sociocultural beliefs, values, and expectations that lie beneath the surface. These less obvious components exert a substantial influence on students' learning engagement and interpretation of educational experiences. For example, implicit biases, unwritten social norms, and cultural attitudes toward achievement often shape student behavior and performance in ways that may not be immediately apparent.

Integrating Explicit and Hidden Curricula into Assessment

To holistically evaluate student performance and achievement, it is crucial to consider both explicit and hidden curricula. Assessments should not only measure surface-level learning outcomes, such as mastery of content and skills but also explore the deeper and more nuanced aspects of the learning environment that influence those outcomes. By addressing both levels, educators can uncover barriers to learning that may stem from cultural disconnects, unspoken classroom norms, or inequitable practices, thus enabling educators to create more supportive and effective educational experiences.

This comprehensive approach acknowledges that the complex interplay between visible and hidden factors shapes student success. Assessments designed with this dual focus not only provide a more accurate picture of student learning but also ensure that all aspects of the educational environment contribute positively to student growth.

Methods for Assessing Student Growth and Performance

The next step is to examine the methods used to assess student growth and performance with a clear understanding of the theoretical foundations and diverse characteristics that influence student learning. Effective assessment methods not only measure what students know but also provide meaningful insights into how they learn, grow, and apply their knowledge over time. This section explores various assessment approaches ranging from formative and summative methods to diagnostic and performance-based techniques. By employing varied and well-aligned assessment methods, educators can gain a more holistic understanding of student achievement, identify areas for improvement, and tailor instructional strategies to support ongoing development. These methods are essential to foster a growth-oriented learning environment that empowers students to achieve their full potential.

Formative Assessments

Formative assessments are powerful tools in education that provide ongoing evaluations to inform instructional decisions and support student progress (Bennett, 2011). By identifying gaps in understanding, they enable timely intervention and ensure that teaching aligns with the evolving needs of the students. Hattie (2009) emphasized the significant impact of effective feedback from formative assessments on student achievement, while Black and Wiliam (1998) demonstrated how continuous, personalized feedback enhances learning outcomes by addressing individual needs.

A practical example can be seen in a fourth-grade classroom where a teacher uses formative assessment to improve students' understanding of multiplication. After administering a quiz, the teacher identified common misconceptions and adjusted subsequent lessons to address these areas of confusion. This process exemplifies how formative assessments provide dynamic real-time insights that guide teaching and learning (Airasian et al. 2012).

Far from mere measurement tools, formative assessments foster a feedback-rich environment in which educators can adapt their instruction to promote mastery and deeper understanding (Sadler, 1989). By empowering students with actionable insights and guiding educators to refine their methods, formative assessments are integral to achieving meaningful student-centered learning outcomes (Bennett, 2011).

Summative Assessments

Summative assessments are a vital part of the educational assessment framework intended to evaluate student learning after an instructional period. These assessments commonly take the form of final exams, standardized tests, or end-of-year projects, providing a snapshot of a student's academic accomplishments at a specific time. Although they do not directly affect current learning processes, summative assessments play a crucial role in offering a comprehensive measure of students' overall academic performance.

According to Black and Wiliam (1998), striking a proper balance between formative and summative assessment methods is essential to enhance the effectiveness of the learning process. This dual approach ensures that while formative assessments provide ongoing feedback and opportunities for immediate improvement, summative assessments provide a comprehensive overview of a student's learning trajectory over an extended period.

A notable instance of summative assessment in an eighth-grade science course not only evaluates students' overall understanding but also serves multiple purposes within the educational framework. This comprehensive evaluation, typically administered at the end of the semester, encompasses a wide range of topics covered throughout the term, including subjects such as biology, chemistry, physics, and earth sciences. The evaluation can be conducted through various methods, such as objective tests, brief written responses, hands-on laboratory experiments, or even project-based assignments that require students to apply their cumulative knowledge to real-world scenarios.

Therefore, the strategic utilization of summative assessments is not merely a means of evaluating the culmination of learning but rather a method of integrating these insights to inform subsequent instructional strategies and curriculum development. These assessments function as vital tools for teachers and decision-makers to gauge the effectiveness of educational initiatives and pinpoint areas for improvement. This process helps ensure that learning goals are achieved and that students are well-prepared for the next steps in their academic and career paths.

Diagnostic Assessments

Diagnostic assessments are essential at the beginning of the learning period. They provide valuable insights into students' strengths and areas of improvement. These assessments are specifically designed to identify gaps in knowledge and skills, thus enabling educators to plan instruction in a targeted and effective manner.

Guskey (2003) emphasized the critical role of diagnostic assessments in uncovering students' prerequisite knowledge and skills. This foundational understanding helps educators design instructional strategies that are both relevant and appropriately challenging, ensuring that students are set up for success.

A clear example of the utility of diagnostic assessments can be found in secondary school algebra courses. At the start of the academic year, an instructor administers a diagnostic assessment to evaluate students' prior understanding of fundamental algebraic concepts. Based on these results, the teacher identified common gaps in knowledge and adjusted the curriculum accordingly. For instance, they may allocate more time to reviewing challenging topics, providing targeted interventions for struggling students, or designing differentiated learning activities to address varying levels of readiness. These insights will enable teachers to create a

responsive learning environment that ensures all students are appropriately supported and challenged.

Diagnostic assessments do not solely measure student capabilities; they are also fundamental tools for improving educational quality. By identifying students' diverse needs early, teachers can develop personalized teaching strategies that optimize learning outcomes and foster an inclusive, engaging environment (Bray & McClaskey, 2015; Felder & Brent, 2005; Hargreaves, 2018). Thus, diagnostic assessments serve as the foundation for a successful and responsive educational experience, ensuring that every student can achieve their full potential.

Ipsative Assessments

Ipsative assessments offer a distinctive and valuable approach to educational evaluation by focusing on students' personal growth and development. Unlike traditional assessments, which compare student performance to external benchmarks or standards, ipsative assessments measure progress by comparing a student's current performance with their past achievements. This shift in focus from comparative achievement to individual improvement emphasizes personal milestones, fostering a deeper sense of accomplishment.

Taras (2005) highlighted the motivational benefits of ipsative assessments, noting that concentrating on individual progress and self-improvement, these assessments can significantly enhance a student's drive to succeed. When students clearly see evidence of growth, they reinforce positive feedback, build confidence, and validate their efforts. This tangible recognition of progress is especially empowering for learners who may struggle in competitive or high-pressure environments, encouraging them to view these challenges as opportunities for further development.

An illustration of an ipsative assessment can be seen in a fifth-grade classroom, where a student maintains a writing portfolio to document her progress over the academic year. The portfolio may include essays written at various points in time, showcasing the evolution of her writing skill. An essay from September might demonstrate basic sentence structures and limited vocabulary, while essays from later in the year reveal more sophisticated sentence construction, expanded vocabulary, and improved organization. This clear evidence of progress fosters a sense of achievement and encourages students to continue progressing. The

reflective process of reviewing earlier work further reinforces her growth, bolstering her self-efficacy and motivation to improve.

Ipsative assessments nurture a growth mindset by encouraging students to view their educational journey as a continuous process of self-improvement rather than a comparison with peers. This approach fosters more engaged and self-motivated learners who recognize and value their progress. By focusing on individual milestones, ipsative assessments not only evaluate learning outcomes but also contribute to creating a positive and empowering learning environment in which students feel supported by their personal growth.

Norm-Referenced Assessments

Norm-referenced assessments are foundational tools in education designed to compare students' personal results with those of a broader reference group. Educators and policymakers rely on comparative data to identify trends, allocate resources, and implement targeted interventions to address learning gaps and improve educational outcomes.

Stiggins (1977, 2001) emphasized the significance of norm-referenced assessments in fostering accountability and providing a macro-level view of student achievement. To illustrate, eleventh-grade history students might be required to take a standardized test administered across the state to evaluate their grasp of significant historical occurrences and principles. The results not only reflect individual performance but also allow comparisons across schools and districts. Such data are instrumental in evaluating curriculum effectiveness, refining instructional strategies, and informing resource allocation.

These assessments are more than mere measurement tools; they are integral in maintaining educational standards and addressing systemic challenges. For instance, if statewide data reveal consistent struggles with specific historical concepts across multiple schools, educators and policymakers may reevaluate instructional approaches or curriculum materials. Additionally, norm-referenced assessments can highlight disparities in educational outcomes among different student groups, prompting further analysis of contributing factors such as socioeconomic or cultural influences (Chappuis & Stiggins, 2016).

By facilitating a comparative understanding of student performance, norm-referenced assessments support efforts to identify the strengths and weaknesses

of an educational system. When thoughtfully used, they can guide decisions to promote equity and enhance the overall quality of education.

Criterion-Referenced Assessments

Criterion-referenced assessments evaluate student performance based on specific criteria or predefined learning standards in contrast to norm-referenced assessments, which measure students' performance relative to their peers (Popham, 2008). This approach fosters a transparent and goal-oriented learning environment, enabling students to work toward specific competencies. These assessments provide an accurate measure of individual progress by assessing performance against predetermined standards. They help teachers recognize areas where students need further assistance or advanced learning opportunities and allow instruction to be customized to effectively address diverse educational needs.

Popham and Husek (1969) emphasized that criterion-referenced assessments primarily outline students' expected knowledge and abilities at various points in their educational progression. These assessments establish clear objective criteria and offer a concrete framework for instruction and evaluation. This clarity benefits educators by helping them target their teaching strategies and provide students with a clear understanding of their expectations.

An example of criterion-referenced assessment is in a third-grade math class where students are evaluated on specific geometric standards such as identifying shapes, understanding angles, and solving geometric problems. Teachers use these results to differentiate instruction and tailor lessons to address their students' individual needs (Bennett, 2011). For instance, students excelling in shape identification might advance to more complex three-dimensional figures. In contrast, those struggling with angles might receive hands-on activities or visual aids to reinforce their understanding. This targeted approach ensures that all students receive appropriate instruction, fostering an inclusive learning environment that meets diverse needs based on performance against defined criteria.

Criterion-referenced assessments are essential tools for educators to evaluate whether students have successfully achieved the competencies and knowledge delineated in their curriculum. Their clarity and specificity help to align teaching strategies with learning objectives, thereby creating a structured and goal-oriented educational environment. Reassessment using the same criteria enables educators to monitor student progress over time, track growth, and enhance their

understanding. Ultimately, these assessments support effective instructional planning, personalized learning, and a focused evaluation approach that fosters successful student development.

Self-Assessments/Reflective Assessments

Self-assessment is a vital component of educational evaluation that engages students in actively evaluating their learning and performance. This practice fosters metacognition and self-regulation, helping students to develop a deeper understanding of their strengths and areas of improvement. Through this heightened awareness, students can set realistic goals and take ownership of their learning journey. Moreover, self-assessment equips students with transferable skills that extend beyond the classroom, thereby supporting lifelong learning and professional growth.

Research has highlighted the significant role of self-assessment in the development of essential skills among students. Zimmerman (2001) emphasized that self-assessment promotes active participation in learning, fostering autonomy and self-direction. By reflecting on their progress, students enhance their self-awareness, enabling them to set achievable goals and monitor their development effectively. This process supports sustained learning and personal growth by encouraging introspection and strategic decision-making.

A practical example of self-assessment in high school literature classrooms demonstrates its ability to drive student growth. When students evaluate their essay-writing skills, they engage in a metacognitive process that promotes critical thinking. This reflection allows them to pinpoint areas for improvement, such as crafting stronger thesis statements, incorporating relevant textual evidence, and improving essay coherence. By identifying these specific needs, students can adopt a more focused and strategic approach to improving their writing.

The integration of self-assessment into educational practice is central to fostering autonomous learning. It cultivates critical thinking and self-analytical skills, empowering students to recognize their strengths and weaknesses. This recognition fuels continuous learning and improvement, thus creating a foundation for lifelong growth. Incorporating self-assessment into the curriculum encourages students to become engaged, reflective, and motivated learners who confidently direct their educational journeys.

Holistic Assessments

Holistic assessments focus on evaluating the overall quality of student work, emphasizing a comprehensive understanding of student abilities rather than dividing performance into separate criteria. This method provides a fuller picture of a student's understanding, critical thinking, and application of knowledge by recognizing the interconnectedness of various skills and concepts. By assessing the entirety of a student's work, educators can gain deeper insights into their strengths, areas for improvement, and overall competencies, thereby offering a more authentic representation of their abilities.

As noted by O'Sullivan and Harris (2004), holistic assessments capture the nuanced and accurate representation of a student's overall potential, which traditional segmented evaluations often fail to achieve. By examining the interplay between different competencies, this approach reveals patterns in student understanding that might otherwise go unnoticed.

In creative subjects such as art, holistic assessments are particularly valuable for evaluating multiple aspects of student work simultaneously. For instance, in a second-grade art class, a teacher might assess a student's painting, considering its creativity, expression, and composition rather than isolating specific technical elements. This method encourages young learners to explore their creativity freely and foster self-expression and confidence (Dewey, 1938). By providing feedback that recognizes individual artistic voices, teachers can nurture students' artistic growth in a developmentally appropriate manner, even if technical skills are still emerging. Holistic assessments also enable educators to tailor their instructions to support each student's unique needs, thus creating a positive learning environment that prioritizes creativity and exploration.

Holistic assessments closely align with real-world applications in which skills and knowledge are seldom used in isolation. This approach encourages students to integrate their learning across subjects and domains, fostering an adaptive and interconnected understanding. Educators benefit from a more comprehensive view of students' critical thinking, problem-solving, and ability to apply knowledge in diverse contexts (Broadfoot et al., 2002). This perspective not only enhances the assessment process but also informs instructional strategies, allowing teachers to support individual student growth and development better.

Curriculum-Based Assessment (CBA)

Curriculum-based assessment (CBA) aligns closely with the curriculum being taught, offering educators a targeted method for evaluating students' skills and knowledge in relation to specific instructional objectives. This alignment ensures that assessments are relevant and meaningful, providing valuable insights into individual student progress and the effectiveness of teaching methods. CBA encompasses a variety of formats, including quizzes, performance tasks, observations, and portfolio evaluations, and is designed to capture a comprehensive picture of student learning (Newmann & Archibald, 1992).

Fuchs and Fuchs (1986) highlight the significance of CBA in guiding instructional adjustments. By aligning assessments with the curriculum, educators can implement timely and targeted changes to their teaching strategies, ensuring that instruction is responsive to student needs. Regular feedback generated through CBA fosters a culture of continuous improvement, enhancing teaching effectiveness and student outcomes.

For example, in a sixth-grade science classroom, the teacher employs CBA through hands-on experiments, data collection, and result presentations that mirror real-world scientific practice. This method assesses not only students' understanding but also their capacity to implement scientific concepts, engage in critical thinking, and effectively convey their results. Teachers can obtain a thorough understanding of students' scientific proficiencies by observing them throughout the scientific process. This includes watching students create hypotheses, constructing experimental designs, examining the collected data, and formulating conclusions based on their findings. Over time, this method tracks student growth in specific areas such as controlling variables, improving measurement precision, and interpreting data. The interactive nature of these assessments also fosters collaboration and peer learning, enriching the educational experience. Moreover, the data collected help the teacher identify areas where individual students or the class may need additional support or enrichment, enabling a more personalized and effective instructional approach.

The CBA's influence goes beyond evaluation and actively contributes to the learning process itself. Immediate feedback allows students to identify their strengths and weaknesses, fostering self-reflection (Brown & Harris, 2013) and metacognition. This constant cycle of input mitigates assessment-related stress,

cultivates an attitude centered on personal development, and encourages learners to see evaluations as opportunities for self-improvement rather than high-stakes judgments. Ultimately, the CBA transforms assessment into a dynamic tool to facilitate and enhance learning, thereby creating a holistic and effective educational experience.

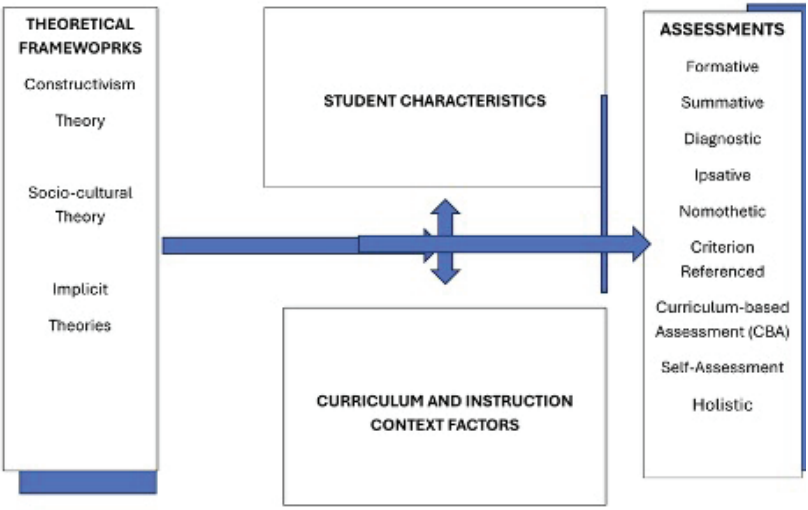
Summary and Conclusion

This chapter presents a theory-informed and student-centered assessment framework, emphasizing not only performance evaluation but also nurturing a positive mindset, recognition of individual differences, and adaptation to diverse educational contexts. By incorporating a spectrum of assessment strategies, educators can create an environment where students feel supported, motivated, and empowered to actively engage in their learning journey, contributing to their holistic development.

A balanced assessment approach that integrates formative, summative, diagnostic, self-assessment, holistic, ipsative, norm-referenced, criterion-referenced, self-reflective, and curriculum-based assessment methods is essential to promote meaningful learning and academic growth. The continuous cycle of formative assessments, along with periodic summative evaluations and targeted diagnostic tools, provides a comprehensive understanding of student progress and supports individualized learning (Brookhart, 2009, 2010). Self-assessment and holistic evaluations further encourage students to evaluate themselves and engage in comprehensive assessments, thereby promoting a sense of responsibility for their education. This approach nurtures metacognitive skills, enhances analytical thinking, and instills lasting dedication to continuous personal development. Thoughtfully applying these strategies empowers students to reach their full potential and fosters a growth mindset while mitigating the risks of fixed mindsets (Dweck, 2006).

Overall, the assessment methods in this chapter serve a unique and essential role in contributing to a comprehensive understanding of students' abilities, promoting academic excellence, and creating a culture of continuous improvement. By thoughtfully implementing diverse assessment methods and strategically incorporating various pedagogical approaches, teachers can cultivate inclusive and productive learning environments that inspire students to attain their full academic potential. Lastly, this chapter emphasizes the critical role of aligning assessment practices with instructional goals and student characteristics to ensure that the educational process is equitable, supportive, and impactful.

Figure 1.
Theory-Informed Student-Centered Assessment Framework



Note: Figure 1 demonstrates a visual representation of the Theory-Informed and Student-Centered Assessment Framework presented in this chapter.

References

Ainscow, M., Booth, T., & Dyson, A. (2006). Inclusion and the standards agenda: Negotiating policy pressures in England. *International Journal of Inclusive Education, 10*(4–5), 295–308. https://doi.org/10.1080/13603110500430633

Airasian, P. W., & Russell, M. K. (2012). *Classroom assessment: Concepts and applications*. McGraw-Hill.

Armour-Thomas, E., & Gordon, E. W. (2012). *Toward an understanding of assessment as a dynamic component of pedagogy.* Educational Testing Service. http://www.gordoncommission.org/rsc/pdf/armour_thomas_gordon_understanding_assesment.pdf

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bereiter, C., & Scardamalia, M. (2012). *What will it mean to be an educated person in the mid-21st century?* Educational Testing Service. https://www.pt.ets.org/Media/Research/pdf/bereiter_scardamalia_what_will_mean_educated_person_century.pdf

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bray, B., & McClaskey, K. (2015). *Make learning personal.* Corwin.

Broadfoot, P., Daugherty, R., Gardner, J., & Gipps, C. (2002). *Assessment for learning: Beyond the black box*. University of Cambridge, School of Education. https://doi.org/10.13140/2.1.2840.1444

Brookhart, S. M. (2009). *Exploring formative assessment.* ASCD Publishers.

Brookhart, S. M. (2010). *Formative assessment strategies for every classroom* (2nd ed.). An ASCD Action Tool Publishers.

Brown, G. T. L., & Harris, L. R. (2013). Student Self-Assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367–393). SAGE. https://doi.org/10.4135/9781452218649.n21

Brown, R., & Miller, L. (2021). Real-time analytics in education: Shaping instructional strategies. *Journal of Learning Analytics, 8*(1), 45–62. https://doi.org/10.18608/jla.2021.7136

Bryant, D. P., Bryant, B. R., & Smith, D. D. (2019). *Teaching students with special needs in inclusive classrooms*. Sage Publications. https://doi.org/10.4135/9781544365015

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83*(1), 1–6. https://doi.org/10.1080/00098650903267784

Chappuis, J., & Stiggins, R. J. (2016). *An introduction to student-involved assessment for learning* (7th ed.). Pearson.

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning: A systematic and critical review*. Learning and Skills Research Centre.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3–4), 325–346. https://doi.org/10.1207/s15326985ep2603&4_6

Dewey, J. (1938). *Experience and education*. Macmillan.

Diaz, R. M., & Berk, L. E. (1992). *Private speech: From social interaction to self-regulation*. Lawrence Erlbaum. https://doi.org/10.4324/9781315807270

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Edmondson, A. (2023). *Right kind of wrong: The science of failing well*. Simon Element; Simon Acumen.

Felder, R. M., & Brent, R. (2005). Understanding student differences. J*ournal of Engineering Education, 94*(1), 57–72. https://doi.org/10.1002/j.2168-9830.2005.tb00829.x

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199–208. https://doi.org/10.1177/001440298605300301

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism, and education.* Palgrave Macmillan.

Gordon, E. W., Aber, J. L., & Berliner, D. (2012). C*hanging paradigms for education: From filling buckets to lighting fires to cultivation of intellective competence.* Educational Testing Service. http://www.gordoncommission.org/rsc/pdf/gordon_gordon_berliner_aber_changing_pardigms_education.pdf

Gordon, E. W., et al. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical report*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Gordon, E. W., & Rajagopalan, K. (2016). The Gordon commission and a vision for the future of assessment in education. In E. W. Gordon & K. Rajagopalan (Eds.), *The testing and learning revolution* (pp. 1–8). Palgrave Macmillan. https://doi.org/10.1057/9781137519962_1

Guskey, T. R. (2003). *How classroom assessments improve learning. Educational Leadership, 60*(5), 6–11.

Hall, E. T. (1976). *Beyond culture.* Anchor Press.

Hambleton, R. K., & Zenisky, A. L. (2011). *Translating and adapting tests for cross-cultural assessments.* Routledge. https://doi.org/10.1017/CBO9780511779381.004

Hargreaves, A. (2018). Personalized learning and the next generation of assessment. *Journal of Educational Change, 19*(4), 367–393. https://doi.org/10.1007/s10833-018-9330-1

Hattie, J. (2009). Visible learning: *A synthesis of over 800 meta-analyses relating to achievement*. Routledge Publishing. https://doi.org/10.4324/9780203887332

Jackson, P. W. (1968). *Life in classrooms*. Holt, Rinehart & Winston.

Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 71–83). State University of New York Press.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students*. Pearson.

Piaget, J. (1952). *The child's conception of number.* Routledge & Kegan Paul.

Piaget, J. (1970). Piaget's theory of cognitive development. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 1, pp. 703–732). Wiley. https://doi.org/10.1002/9781119171492.wecad364

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33

Popham, W. J. (2008). *Transformative assessment.* Association of Supervision and Curriculum Development. https://www.ascd.org/books/transformative-assessment

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9. https://doi.org/10.1111/j.1745-3984.1969.tb00654.x

Reeves, D. B. (2017). *The transformational power of student teams: A guide for teachers in a professional learning community.* Solution Tree Press. https://www.solutiontree.com/the-transformational-power of-student-teams.html

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2),119–144. http://dx.doi.org/10.1007/BF00117714

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27. https://doi.org/10.3102/0013189X021004022

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Snow, R. E. (1977). Individual differences and instructional theory. *Educational Researcher, 6*, 11–15. https://doi.org/10.3102/0013189X006010011

Solano Flores, G. (2016). *Assessing English language learners: Theory and practice* (1st ed.). Routledge. https://doi.org/10.4324/9780203521953

Stiggins, R. J. (1977). An introduction to student-involved assessment for learning. *Educational Assessment, 4*(1), 1–14. https://rickstiggins.com/book/introduction-to-student-involved assessment-for-learning/

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Merrill Prentice Hall.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive Science, 12*, 257–285. https://doi.org/10.1207/s15516709cog1202_4

Valdés, G. (2018). Analyzing the curricularization of language in two-way immersion education: Restating two cautionary notes. *Bilingual Research Journal, 41*(3), 1–25. https://doi.org/10.1080/15235882.2018.1539886

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://www.jstor.org/stable/j.ctvjf9vz4

Vygotsky, L. S. (1986). *Thought and language.* MIT Press.

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky, Volume 1: Problems of general psychology* (*pp. 39–285*). Plenum Press. (Original work published 1934.) https://www.marxists.org/archive/vygotsky/works/words/Thinking-and-Speech.pdf

Weiner, B. (1974). *Achievement motivation and attribution theory.* General Learning Press.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research*. SAGE Publications. https://doi.org/10.4324/9780203772218

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 1–37). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410601032

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29*(3), 663–676. https://doi.org/10.3102/00028312029003663

Validity for Assessments Intended to Serve Learners

Stephen G. Sireci and Danielle Crabtree

Abstract

Validity has been described as the most important characteristic in educational testing; yet many practitioners are unfamiliar with the terminology and what needs to be done to ensure tests are fulfilling their intended purposes. In this chapter, we define validity and describe the importance of validating educational assessments designed to serve learners. We discuss how a focus on validity and the practice of validation can help assessments in the service of learning accomplish their goals. Such validation involves collaborating with educators, parents, guardians, and students to design assessments that provide value to them and support learning. We point out such collaborations are typically not incorporated into test design and validation, but are essential for educational assessments to reach their potential in serving learners.

Validity for Assessments Intended to Serve Learners

There are many associations of the word *validity*. One might think of a valid argument, which is one that is well reasoned and conclusive because it is based on sound logic and evidence. Others might think of the legitimacy of a certain claim, or of the facts supporting a point someone made in a dispute. These understandings of validity are similar to how the term is used in educational testing. Given that tests are developed and used for specific purposes, the degree to which those purposes are fulfilled is important to evaluate and document. Equally important is evaluating and documenting the degree to which there are unintended side effects associated with the testing process.

Educational assessments designed to serve learners have a common, general purpose to support student learning. Examples of more specific purposes include engaging students in their learning, providing information to students and their teachers about how they learn, reflecting students' strengths, building their academic self-efficacy, and indicating areas in need of improvement. The goals of assessment in the service of learning are consistent with the goals of instruction in the service of learning. Hence, these "educative assessments" can be thought of as pedagogical tools. However, for assessments to truly serve learners they must not only demonstrate evidence of such service, they must abide by *Principles of Assessment in the Service of Learning* (Baker et al., 2025; Volume I of this *Handbook* series) that require assessment practices to be transparent to stakeholders, grounded in high-quality evidence, and geared toward feedback that improves teaching and learning.

Many of the chapters in this *Handbook* describe theories and methods for designing and conducting assessment in the service of learning. One may ask, "Do these theories and methods work?" "Do these assessments actually serve learners?" These evaluative questions are the focus of the present chapter because in educational testing, investigations of "validity" encompass the evaluation of the effectiveness, utility, and consequences of an assessment. In this chapter, we first define validity and the process of *validation*. We then discuss how a focus on validity and the practice of validation can help assessments in the service of learning conspicuously accomplish their goals. Elevating the voices of educators, parents, guardians, and students is a key feature in designing and validating assessments in the service of learning. We believe these voices have been left out

of many prior validation efforts, which has seriously undermined the degree to which educational assessments have reached their potential to serve learners.

What is Validity?

In educational assessment, *validity refers to the degree to which evidence and theory support the use of a test for a particular purpose.* Evidence refers to the results from strategic and comprehensive research designed to evaluate how well a test is fulfilling its intended purposes and avoiding any unintended, harmful consequences. Theory refers to the conceptualizations describing what the test is measuring, and the logic describing how the interpretations and use of test results will fulfill the intended purposes.

This definition implies several attributes of validity in educational testing. Specifically,

- Validity is not a characteristic of a test, but rather a characteristic of the interpretations, actions, and outcomes associated with a testing process.
- The purposes and intended uses of a test must be clearly specified, because they communicate the testing claims to stakeholders, and set the targets for validation.
- If test results are to be used for multiple purposes, each use must be supported by validity evidence.

Note that the second and third bullets build transparency in testing through a shared understanding between those who create or commission the tests, and those who use them. That shared understanding relates to the purpose of the assessment, the specific knowledge and skills being tested, and the intended use of test results and how such use will support learning.

Validity is an overarching standard against which all tests are evaluated. In addition to the attributes described thus far, it is also important to note validity evidence must include both evidence confirming the appropriateness of the use of the test for its intended purposes and *evidence that use of the test does not result in unintended harmful consequences*. This latter point is especially important for assessments in the service of learning because their explicit purposes are to

evoke only beneficial consequences such as increased learning, increased love of learning, and affirming the learners' enormous capacity to learn.

Validity is sometimes quickly joined to another component of assessment quality—reliability. Reliability refers to the consistency of test scores, that is, if students were tested over and over again, would they get the same test scores—just as if we step on a bathroom scale over and over again we would get the same reading of our weight. Reliability is related to validity in that if scores from an assessment are not reliable, that means students can get very different scores because the test itself or the testing process is unstable. Thus, consistency in the information provided by an assessment (i.e., reliability) is a prerequisite for validity (for more on reliability, see Bandalos, 2018).

Before further describing validity with respect to tests intended to support learners, we discuss the process evaluating educational tests, which is called *validation,* or a *validity investigation*.

What is Validation?

Validation is the process of evaluating the validity of the use of a test for a particular purpose. It involves carefully planned and ongoing research for critiquing the theory underlying test development and for gathering validity evidence. Validation research begins with the earliest stages of test development (e.g., defining the testing purpose and the targets of measurement[1]) and continues throughout all phases of the lifecycle of a testing program. The goal of validation is to provide a sound basis for interpreting and using test scores and other information resulting from the testing process. In addition, validation research can be, and should be, used to improve tests and testing processes whenever the evidence suggests improvements are warranted.

In a subsequent section we return to the topic of validation to indicate how evidence can be gathered to evaluate the degree to which a test designed to support learners accomplishes its intended goals. First, however, we discuss how those goals should be articulated, and include some examples of specific purposes of assessments in the service of learning

1 As we subsequently describe, the targets of measurement are often referred to as "constructs," because theories are *constructed* about the unobservable processes that explain test takers' thoughts and behaviors when responding to test items (Sireci, 2020).

Goals, Purposes, and Uses of Assessments in the Service of Learning

The common goal of all assessments for the service of learning is to benefit the learner. Learners can gain direct benefits and indirect benefits, depending on the purpose and use of the assessment. Direct benefits are those the learner experiences before, during, and after the administration of the test. Before the assessment, the learner may take the time they need to study, to inform themselves of the assessment's construct(s), and to hold themselves accountable for their own learning. In this way, the assessment is viewed as an opportunity for development and becomes a crucial part of the learning process. During administration, the learner may benefit from the assessment metacognitively, gaining instantaneous insight into the aspects of the construct(s) they know well, the aspects that still remain uncertain, and the aspects they have yet to learn. After administration, learners gain the benefits of feedback and reflection, allowing them to self-evaluate their learning, to continue to grapple with the concepts they are still trying to comprehend, and providing them control over their own future learning. However, for learners to reap the benefits of reflection, feedback should include free access to the assessment, time to process how the assessment shaped their learning, and clarifying support for confusing concepts.

Indirect benefits are those learners experience due to the assessment-based decisions made by other members of the school community. A common example is a diagnostic assessment triggering an intervention for a learner. In this case, the learner is not necessarily part of the decision-making process (though, they could be), but improvement to their learning is the focal point for all decisions being made. Effective assessments for the service of learning allow educators to pinpoint the learner's misconceptions, provide information that enables the educator to develop personalized instruction to fit the learner's needs, and allow for progress monitoring, to ensure the skills the learner is supposedly developing are actually being developed. Thus, a first step in ensuring and evaluating the validity of an assessment in the service of learning is to clearly state the direct and indirect benefits. These benefits are the goals of the assessment process and so in validation, we are required to look for evidence that they are being realized.

Although the overall goal is to benefit the learner, the purpose of the assessments in the service of learning is to ascertain specific and insightful information into the learner's educational growth and progress. The type of information obtained and

the depth of insight is dependent on the type of assessment used to gather the information. An extensive list of common and emerging models for assessments in the service of learning have been compiled by Gordon (2020). Each model promotes its own purpose so it is important that the purpose for any assessment is transparent and explicitly stated for all those involved in the testing. Therefore, validation of the assessment requires evidence that the stated purpose is in alignment with its intended goals, use, and consequences.

The basic use of assessments in the service of learning is to inform instruction and learning. Assessments as pedagogical tools act as catalysts between instruction and learning, providing the momentum necessary for educational growth. Yet, it is important to remember assessments are just snapshots of time, measuring the learner's knowledge at that moment for a specific purpose. The purpose, administration, and timing of these snapshots are important. Traditionally, assessments used before instruction measure prior knowledge, assessments used during instruction measure the learner's current understanding, and assessments used after instruction are used to measure retention. These traditional approaches to assessment are evaluative, closed-book, and are used to measure "the beginning" or "the end" of learning. Yet, learning is continuous and so the assessments meant to propel learning forward must also be continuous. That is not to say that learners need to be caught in a constant cycle of evaluation. Rather, assessments in the service of learning are used to promote continuous reflection, allowing one to assess what has been learned and what still needs to be learned at a time and in a way that is most informative. This means the purpose, administration, and timing of the assessment is done in collaboration between the learner and those who are providing the instruction.

Moreover, the use of assessments in the service of learning involves planned intended consequences. If an assessment is used to inform instructional decisions, then the consequences of those instructional decisions must also be considered and stated explicitly prior to the assessment. To be clear, consequences in this circumstance do not imply positive or negative outcomes (which are subjective), but rather the consequences that are necessary to benefit the learner. Additionally, planned consequences do not only pertain to the learner. For example, a change in instructor may be deemed an appropriate consequence that would benefit the learner. Once again, planning intended consequences should be a collaborative effort, ensuring all stakeholders are in agreement that the consequences align

with the appropriate instructional decisions, and are for the learner's best interest, as supported by the evidence collected through validation. As clearly stated in the *Principles for Assessment in the Service of Learning* (Baker et al., 2025), helpful and supportive feedback is a critical intended consequence for assessments to serve learners. The next steps suggested by feedback should help support the intended consequence of learning, including the socioemotional aspects such as improved academic self concept.

Validating Assessments in the Service of Learning

Our prior discussion of how we can articulate the purposes of specific assessments designed to serve learning has set us up for discussing how we can evaluate their intended and unintended consequences. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014) specifies five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). These five sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) testing consequences. We believe these five sources are helpful in evaluating assessments in the service of learning, but we add an additional source of validity evidence that comes from engaging with traditionally forgotten stakeholders in the assessment process—teachers, students, parents, and guardians. Borrowing from O'Leary et al. (2017), we refer to this source of validity evidence as validity evidence based on analysis of test score interpretations.[2]

Sources of Validity Evidence

The word "evidence" has been used numerous times in our discussion of validity and test validation. In this section, we describe specific sources of validity evidence that are helpful for building a validity argument and for demonstrating assessments are of sufficient quality for fulfilling their intended purposes. These sources have been described in the past as different types of validity, but characterizing them in that way is a misnomer. As described earlier, validity refers to the degree to which evidence and theory support the use of a test for a particular purpose. Thus, we cannot divide this concept into types or subtypes. However, we can distinguish

2 At the time of this writing, the AERA et al. Standards are under revision and it is likely this sixth source of validity evidence may be added into the next version.

among the different sources of evidence that can be used to evaluate validity, and represent important aspects of the quality of an educational assessment.

What follows are general descriptions of six sources of evidence that may be useful for evaluating the degree to which assessments designed to serve learners fulfill their goals and do not lead to unintended, negative consequences. These descriptions are intended to describe specific sources of validity evidence that have historically been used to develop validity arguments for testing programs. However, they are not considered representative of all potential sources, and one or more sources may not be applicable to a particular testing situation.

Validity evidence based on test content.
Validity evidence based on test content is used to confirm the content of the test is representative of the knowledge and skill domain the test is designed to measure and is consistent with the testing purpose. The content of a test includes directions, items (tasks), and the stimuli associated with items. In educational testing, validity evidence based on test content is most often gathered by using subject matter experts to review the targets of instruction (e.g., curriculum standards, objectives associated with a lesson, etc.) and to review the test specifications and test items. The subject matter experts should be external to the test development process so their appraisals can be considered independent and unbiased. These experts are trained, and asked to rate test items regarding their degree of alignment to intended objectives, to provide input regarding the degree to which the test represents the targeted knowledge and skill domain, and to evaluate the degree to which students' responses to test items will provide the information intended within the testing purpose.

There are many methods for structuring experts' reviews of tests and items. Traditional methods based on content validity indices have been reviewed by Crocker et al. (1989) and Sireci (1998). These methods typically require experts to indicate the content areas and cognitive levels measured by test items, or to rate how well test items measure specific objectives. Newer methods based on test-curriculum "alignment" evaluate the link between curriculum frameworks, testing, and instruction (Bhola et al., 2003; Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014).

Validity evidence based on test content is particularly relevant to assessment in the service of learning because these assessments should specifically target instructional goals. It is important teachers see the tasks and items that comprise an assessment as reflecting their goals. The tasks presented to learners on these assessments represent the content to be independently evaluated by subject matter experts, who should be experienced with the goals of instruction, and different from the authors of the assessment. These external subject matter experts need to be trained to provide the types of judgments most helpful for determining assessment-instruction alignment. For example, experts can rate the degree to which assessment tasks appropriately measure specific content or cognitive objectives, or capture common misconceptions. An example of a rating task presented to subject matter experts is presented in Figure 1. In this rating task, experts were informed of the content standard (College and Career Readiness Standard for Adult Education—CCRSAE) and cognitive level targeted by each assessment item. They were asked to review each item and complete the four ratings illustrated in Figure 1. These data can be aggregated across raters and across items to provide overall information regarding the content quality of the assessment with respect to instructional alignment. It should be noted the four questions illustrated in Figure 1 are just one example of how to gather data from subject matter experts regarding content quality, and other methods can be used that involve more open-ended questioning, or targeting larger grain size elements of knowledge and skills such as a lesson or curricular unit.

Figure 1.

Example of Item-Level Ratings Gathered to Provide Validity Evidence Based on Test Content

	Not well at all	Slightly well	Moderately well	Very well	Extremely well
How well does the item measure the CCRSAE Standard?	○	○	○	○	○
How well does the item measure the cognitive level?	○	○	○	○	○
How well does the item meet your standards for a high quality item?	○	○	○	○	○
How well do the distractors (incorrect response options) represent realistic mistakes?	○	○	○	○	○

Note: CCRSAE=College and Career Readiness Standards for Adult Education (Pimentel, 2013).

Other evaluation tasks could be presented to subject matter experts to have them gauge the degree to which the tasks presented on the assessment meet the test development targets and are likely to provide the intended information regarding student learning. The degree to which the assessment represents the instructional goals (content representation) and to which all items are relevant to those goals (content relevance) are important evaluation criteria for which data should be gathered. Many methods exist for gathering subject matter experts' perceptions of test content including those based on traditional indices of content validity (e.g., Crocker et al., 1989; Sireci & Faulkner-Bond, 2014) as well as those based on assessment-curriculum alignment (Bhola et al., 2003; Martone & Sireci, 2009). Minimally, teachers and other test developers should get outside opinions from peers regarding the content quality of their assessments intended to serve learners.

One further area in which validity evidence based on test content can be extremely important for assessments in the service of learning is ensuring the content of the test is "culturally responsive," meaning "assessments that take into account the background characteristics of the students; their beliefs, values, and ethics; their lived experiences; and everything that affects how they learn and behave and communicate" (Walker et al., 2023, p. 1). Interrogating test content with respect to the degree to which it is appropriate for the diversity of learners who will interact with the assessment can be conducted using subject matter experts who come from diverse backgrounds and are familiar with the various cultures from which the learners will draw from when interacting with test content. However, culturally responsive assessment is an emerging concept and likely requires more than a finite group of diverse subject matter experts; it will require engaging with test takers themselves and the communities within which they operate (Sireci et al., 2025; Walker et al., 2023).

Validity evidence based on response processes.

Validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers" (AERA et al., 2014, p. 15). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time data, and evaluation of the reasoning processes students use when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009). Such evidence is helpful for confirming assessments in the service of learning are measuring the cognitive skills they intend to measure, and that students are using the targeted skills to respond to the test items.

Given that the overall goal of assessments in the service of learning is to benefit the learner, assessment must go beyond confirming what a learner "knows" or "does not yet know" and provide information on learners' cognitive processes in responding to test items. For example, if a learner consistently answers questions correctly regarding a specific learning objective, we assume the learner has shown proficiency with respect to that objective. However, as educators are well aware, very little if any insight about learning can be gained from a correct answer. Assessments in the service of learning should be designed to probe

deeper into the learner's thinking, not only allowing them to reflect on their own understanding of the construct, but to also see if they can verbalize their thinking and essentially instruct others about the topic. Observations, cognitive interviewing, and open-ended questions are just a few ways to collect this type of information. These qualitative data collected from the assessment, in combination with the quantitative data, provide the evidence needed to make informed decisions regarding instruction and how to further a learner's learning.

Given the importance of validity evidence based on response processes for assessments in the service of learning, we recommend the use of think-aloud and other cognitive probes (See Padilla & Benitez, 2014) to engage learners in the design and revision of these assessments. This participatory approach will reveal how learners perceive what they are asked to do to solve the problems presented on an assessment, what thinking processes they use to solve them, and what misconceptions they may have. The data gathered from these studies can then be compared to the goals of the assessments. When learners confirm the processes they used to solve items are consistent with the same ones the assessment was designed to target, this evidence supports use of the test for its intended purpose.

In addition to direct measures of learners' thought processes when solving test items, log data from digital assessments can be used as an indirect measure of their cognitive behaviors. These log data include the amount of time learners took to respond to an item, what help features they used, the order in which they completed steps to solve a problem, whether they were engaged in trying to solve the item, and whether they skipped or returned to items (Araneda et al., 2022; He et al., 2023; Wise & Kong, 2005).

Validity evidence based on internal structure.

Validity evidence based on internal structure refers to statistical analyses of students' responses to test items. For example, if the items on a test are all designed to measure a single concept or skill, statistical analysis of students' responses should reflect a single "dimension" or "factor" being measured. On the other hand, if an assessment is designed to provide information about multiple skills, then statistical analysis of students' response data should reveal separate, multiple dimensions. Procedures for gathering validity evidence based on internal structure are technical from a statistical perspective and include methods such as "factor analysis" (both exploratory and confirmatory) and "multi-dimensional

scaling." Internal structure evidence also evaluates the "strength" or "salience" of the major dimensions underlying an assessment and so would also include indices of measurement precision such as reliability estimates, decision consistency (i.e., consistency of decisions or classifications made if the students were retested), and other measures of precision (e.g., standard errors of measurement, test information, etc.).

Validity evidence based on internal structure can also be used to evaluate if the test is operating in the same way across different types of students. Analyses of potential item bias across different types of students (i.e., differential item functioning) or of differences in dimensionality across different types of students (e.g., invariance of test structure) also fall under the category of internal structure validity evidence. Rios and Wells (2014) provide excellent examples of conducting such "invariance" of structure analyses.

Validity evidence based on test structure requires a large amount of student response data and so is less applicable to classroom assessment data. In addition, the structure of an assessment will likely differ based on testing students who have or have not yet had the opportunity to learn the material tested. However, if commercial tests are used and administered to students who have already experienced the intended instruction, such evidence should be provided to ensure the types of information provided by the assessment are supported by the dimensionality and precision analyses.

Validity evidence based on relations to other variables.
Validity evidence based on relations to other variables refers to analyses that involve students' test scores as one variable in the analysis along with other variables designed to (a) confirm or disconfirm the test is measuring what it intends to measure, or (b) determine the degree to which test scores are predictive of certain outcomes of interest. An example of confirmatory evidence is the degree to which reading test scores correlate positively with teachers' ratings of students' reading proficiency. In this example, students' test scores are one variable in the analysis and teachers' ratings are the other variable. As an example of disconfirming evidence, a study could be done to evaluate the relationship between students' reading test scores and a measure of test anxiety. The validity of the assessment would be supported if the test scores showed no or little relationship to anxiety. An example of prediction would be studying the degree to which students'

test scores at the midterm of a semester predict their performance on an Advanced Placement exam at the end of the semester. Validity evidence based on relations to other variables typically use the statistical techniques of correlation and multiple regression analyses to demonstrate and evaluate the strength of relationships between test scores and other variables (See Bandalos, 2018; and Sireci & Benitez, 2023, for examples).

In considering how this type of validity evidence can improve or confirm assessment in the service of learning, other indicators of student learning need to be gathered or available. Examples of relevant external variables include the quality and amount of instruction provided; or the amount of time learners interact with assignments, educational games, or other learning tools. Different groups of learners, or different time periods, could also be used as the "other" variable in these analyses. Investigating how test scores differ across students with different learning styles is likely to provide informative validity evidence. Studying relations of other variables to test scores requires test developers, educators, and researchers to be creative in imagining how evidence of student learning could be reflected outside of the assessment context, gathering that evidence, and then analyzing its relationships with test scores.

Validity evidence based on analysis of score interpretations.
Educational tests are developed with explicit intended interpretations and uses of test results in mind. Thus, those developing the tests and those using test results should have a shared understanding of what is represented by a test score, how that score should be interpreted, and what types of actions it can inform. Thus, the actual interpretations made on the basis of test scores should be the same as those that motivated test development. Thus, evidence that test scores are appropriately interpreted by those who use them is a key component of a validity investigation. This source of evidence evaluates the congruence between intended interpretation and uses of scores and actual interpretations and uses.

Gathering validity evidence based on the appropriateness of score interpretation and uses requires that first, these interpretations and uses are clearly defined, and then investigators engage with teachers, students, and other users of test results to study the degree to which they are interpreting test results as intended. Such research is recommended at both the end of a testing process to confirm the intended interpretations are realized, and also early in the test development process

when designing and developing test score reports. By understanding the aspects of test reports that are confusing to or misinterpreted by teachers, students, and other stakeholders, improvements can be made to the communication of test results to facilitate valid interpretation.

Validity evidence based on the consequences of testing.

Evidence based on the consequences of testing refers to evaluation of the intended and unintended consequences associated with a testing program. Intended consequences of testing originate from the explicit use and intended interpretations of the test scores established by the test developer. For example, an intended consequence of a test used for the service of learning may be the placement of a learner into a proper instructional group. Evidence based on this consequence would show support that the test accurately placed the learner into an instructional group that promotes the learner's academic growth and potential. Unintended consequences of testing can originate from various sources, such as culturally irrelevant test items and/or claims that extend beyond the intended interpretations and use of the test scores. For example, a test intended to place a learner into a proper instructional group may have some items that are unintentionally culturally irrelevant to the learner (Solano-Flores, 2019; Randall, 2021). In this case, evidence based on this consequence may show misalignment between learners' true abilities, their performance on the test, and their achievement in the instructional group. Extending this example, the learner's performance on the placement test could also lead to the misinterpretation of the learner's interests or motivation. Evidence based on this consequence may show a negative association between the learners achievement in the instructional group and their actual interest or motivation in what is being taught.

It is important that the evidence based on consequences reflects the reality of the testing situation. Prior experiences with testing can inform ways that potential consequences can be anticipated and addressed. Evidence based on the consequences of testing are going to be unique for every testing situation. However, it is important that the intended and unintended consequences associated with testing are thoroughly evaluated to ensure the learner does not experience any harmful consequences and only receives consequences that are beneficial to their learning.

There are many ways to gather validity evidence based on testing consequences. In some contexts where the results of testing lead to personal rewards such as selection into a competitive school or class, or a credential, the degree to which the test-based rewards differ across demographic groups will be of interest. Such "adverse impact" studies (Sireci & Geisinger, 1998) evaluate the selection or passing rates across groups and compare them to expected, equitable outcomes. Lane (2014) described other studies that can be done to evaluate the consequences of educational tests such as analysis of classroom artifacts, changes in teaching practices, and surveys and interviews of teachers and students.

Synthesizing and Documenting Validity Evidence

Educators who develop assessments for their students may not feel the need to demonstrate a strong body of evidence to support their use, but it is important to provide such evidence to confirm their assessments are doing what they intend to do and demonstrate a level of quality commensurate with their intended uses. Only with such evidence can educators feel confident in the information provided and the decisions they make based on that evidence. For tests and assessment systems produced on a larger scale, clear and sufficient evidence is needed to justify use of the test. In these cases, the body of evidence and theory that support the use of a test for a particular purpose is called a *validity argument* (AERA et al., 2014; Kane, 1992, 2006, 2013). The word "argument" is not intended to infer a fight or a heated debate, but rather a compelling and cohesive synthesis of theoretical analysis and research results that can be used to justify or dismiss the use of a test for a particular purpose. Thus, a validity argument clearly states the purposes and goals of a testing program and provides an informative summary of the research-based evidence gathered to evaluate how well the test (a) measures the intended constructs, (b) provides the information required to fulfill the testing purposes, (c) promotes accurate interpretations of that information, and (d) does not result in harmful unanticipated consequences for individuals or society.

One may wonder when the evidence in a validity argument is sufficient to justify the use of an assessment in the service of learning. Although it will always be a judgment call, the goal is to provide a compelling argument that would satisfy even a reluctant critic. The AERA et al. (2014) *Standards* acknowledge this subjectivity, but note, "at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible" (p. 22).

Concluding Remarks: Valid Assessments in the Service of Learning

In this chapter, we presented a basic overview of the concept of validity in educational testing, the process of test validation, and how they relate to assessments in the service of learning. There is much written about the theory of validity and the process of test validation, and we have provided many references to further reading in those areas. What we believe is most important to keep in mind with respect to the validity and validation of assessment in the service of learning is (a) confirming the assessments are appropriate and beneficial for the learners being assessed, (b) ensuring the assessment model and content represents and measures the intended learning objectives, (c) the assessment properly informs instructional decisions (and the consequences associated with those decisions), and (d) outcomes of the assessment are interpreted as intended, provide valuable insight into the learner's learning processes, and do not result in any undesirable harmful consequences.

With respect to (a) the degree to which assessments in the service of learning are culturally responsive will deserve specific attention. The learners we assess are diverse with respect to culture, language, socioeconomic status, gender identity, neurodiversity, and many other factors. Valid assessments that support learning should provide both mirrors into learners' own cultures as well as "windows" into other cultural groups (Randall et al., 2022).

The other goals of valid assessment can be summarized by stating validity evidence should be provided that assessments in the service of learning appropriately and sufficiently represent their intended learning goals, and do not unintentionally measure other factors. By vigilantly validating our assessments, we not only uphold their effectiveness, but also fulfill the promise of assessment in the service of learning–making our assessment practices more transparent, more high-quality, and more directly informative for every learner's next steps.

We hope the descriptions of validity we provided in this chapter, as well as the examples of validity evidence that can be gathered to evaluate an assessment, are helpful to those who strive to use assessment in the service of learning to help learners learn. Like all endeavors in assessment in the service of learning, facilitating and evaluating validity is a journey in which we continually learn about how we can improve assessments and the assessment validation process. As assessments in the service of learning evolve, we anticipate more examples of validation studies that provide new insights into how we can best use assessments to support student learning.

References

Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring relationships among test takers' behaviors and performance using response process data. *Education Sciences, 12*, 104. https://doi.org/10.3390/educsci12020104

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

Bandalos, D. L. (2018). *Measurement theory and application for the social sciences.* Guilford Press.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21–29.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–194.

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

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

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning, 39*(3), 719–736. https://doi.org/10.1111/jcal.12748

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*,527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp.. 17–64). American Council on Education/Praeger.

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

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction,*Review of Educational Research, 79*(4), 1332–1361.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 13–100). American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 83–108). Charlotte, NC: Information Age Publishing Inc.

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36*(2), 16–23.

Padilla, J., & Benítez, I, (2014). Validity evidence based on response processes. *Psicothema*, 26, 136–144.

Pimentel, S. (2013). *College and career readiness standards for adult education.* U.S. Department of Education, Office of Vocational and Adult Education. http://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: toward a justice-oriented, antiracist validity framework. *Educational Assessment*, https://doi.org/10.1080/10627197.2022.2042682

Rios, J., & Wells, C. S. (2014). Validity evidence based on internal structure. *Psicothema, 26*(*1*), 108–116.

Sireci, S. G. (2020). "De-"Constructing" Test Validation," *Chinese/English Journal of Educational Measurement and Evaluation, 1.* 教育测量与评估双语季刊: https://www.ce-jeme.org/journal/vol1/iss1/3

Sireci, S. G., & Benitez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217–226. https://www.psicothema.com/pdf/4805.pdf

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321.

Sireci, S. G., Crespo Cruz, E., Suárez-Álvarez, J., & Rodriguez, G. (2025). Understanding UNDERSTANDardization research. In *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (R. Bennett, L. Darling-Hammond, & A. Badrinarayan, Eds., pp. 415–433). Routledge.

Sireci, S. G., & Faulkner-Bond, K. (2014). Validity evidence based on test content. *Psicothema.* https://doi.org/10.7334/psicothema2013.256

Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J. H. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.), *Test interpretation and diversity* (pp. 105–140). American Psychological Association: Washington, D.C.

Solano-Flores G (2019). *Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. Frontiers in Education* 4:43. https://doi.org/10.3389/feduc.2019.00043

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally responsive assessment: Provisional principles. *ETS Research Report Series, 2023*(1), 1–24.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Social Justice in Educational Assessment: A Blueprint for the Future

Stephen G. Sireci, Sergio Araneda, and Kimberly McIntee

Abstract

Social justice is a popular topic in education, but it is rarely addressed in educational assessment. Historical attempts to engage the educational assessment community in issues of social justice include several pioneering researchers who pointed out the potential and realized adverse societal effects of testing, as well as the *Standards for Educational and Psychological Testing's* validity evidence based on testing consequences (American Educational Research Association et al., 2014). Drawing from these sources, and from criticisms of the negative effects educational tests have had on individuals and society, we identify several actions that can be taken to promote more socially just testing practices. These actions begin at the earliest stages of test development and extend to test administration, scoring, score reporting, and validation. By considering issues of social justice in test development, we can design assessment systems that better serve education and society, and help end oppressive practices in education, which are central to the goals of assessments that serve learners.

Social Justice in Educational Assessment: A Blueprint for the Future

> *"I don't want no peace. I need equal rights and justice."*
>
> —Peter Tosh

According to Merriam-Webster's online dictionary, the word "just" has several definitions. When used with respect to the social concept of *justice*, the definitions provided are "acting or being in conformity with what is morally upright or good" and "being what is merited." These definitions begin to portray what is meant by the term "social justice," a term first credited to Luigi Taparelli (1840) who defined it as the ability to "maximize individual freedom to associate at all levels" (cited in Boyles et al., 2009; p. 32). This definition was based on Taparelli's belief that smaller communities in society should work together for the common good of larger communities. Today, we may think of Taparelli's vision as a system of "grass roots" efforts; however, the concept of social justice in education has taken on deeper and multifaceted meanings (Porfilio, Strom, & Lupinacci, 2019). In fact, as Boyles et al. pointed out,

> *there are groups promoting education reform in order to perpetuate status quo norms of power and privilege acting in the name of social justice. Yet, and at the same time, there are other groups who wish to dismantle such a privilege under the auspices of social justice.* (p. 30)

For most educators and educational researchers, social justice entails pursuing equity and access to high-quality education for all students regardless of race, sexual orientation, religion, age, and other sociopolitical or sociological characteristics. Although many researchers argue the concept of "social justice" in education cannot be defined in one way (Jean-Marie, Normore, & Brooks, 2009), in discussing how to teach about social justice in education, Bell (1997) characterized social justice education as both a goal and a process. As she described, "The goal of social justice education is full and equal participation of all groups in a society that is mutually shaped to meet their needs…[while]… the process for attaining the goal of social justice… should be democratic and participatory, inclusive and affirming of human agency and human capacities for working collaboratively to create change" (pp. 3–4). In addition to participation as both a goal and a process, common themes across different conceptualizations of social justice in education

include confronting issues of racism, sexism, homophobia, xenophobia, and other manifestations of oppression to create and maintain equitable schooling so all students (broadly defined) can achieve their educational and occupational goals (e.g., Boyles et al., 2009; Jean-Marie et al., 2009; Porfilio et al., 2019).

Although considerable attention has been devoted to social justice in education, the concept has received little attention in the educational testing community. This lack of attention is unfortunate, in that justice is a critical component for educational assessments to serve learners. Our chapter uses themes of social justice to advance key principles of assessment in the service of learning—namely, making assessment processes transparent and fair (Principles 1 and 6) and ensuring high-quality, credible uses of assessments (Principle 7, see Baker et al., 2025). In this chapter, we argue social justice in educational assessment should be front-of-mind for the entire measurement community; and by properly addressing issues of social justice, we can eliminate the contribution of educational tests to systemically racist and unjust educational practices.

Educational tests are an integral and enduring part of contemporary society. They have enormous consequences that may affect the quality of education an individual may receive and the degree to which individuals are able to reach their aspirations. Although there have been criticisms against the widespread use of tests (e.g., Bertrand & Marsh, 2021a, 2021b; Koljatic et al., 2021a), and some measurement scholars have discussed the importance of values and consequences in testing (e.g., Messick, 1989; Shepard, 1993; Sireci, 2021), there has not yet been a comprehensive discussion of social justice in educational assessment. Thus, the purpose of this chapter is to end this deficit in the psychometric literature by (a) tracing the history of concerns for social justice in educational assessment, and (b) proposing actions for more socially just educational assessment practices.

To accomplish these purposes, we first review the guidance provided by professional testing standards. Next, we review seminal literature written by educational measurement researchers and practitioners who addressed issues of social justice. Drawing from our review, we recommend several steps to promote testing practices that are consistent with and useful for an educational system with social justice at its core.

Concerns for Social Justice in Educational Assessment: A Brief History

> *It was not until I was long out of school and indeed after the (first) World War that there came the hurried use of the new technique of psychological tests, which were quickly adjusted so as to put black folk absolutely beyond the possibility of civilization (W.E.B. DuBois, 1940, as quoted in Guthrie,* 1998, p. 55).

As Sireci and Randall (2021) described, the history of contemporary educational testing is often traced back to the early work of Alfred Binet, who developed tests to identify children in Paris in need of special education. Although this purpose was laudable and led to the inclusion of children in schools who otherwise would have been denied an education, there is also a darker side to the history of educational testing. It is this darker side that is reflected in W.E.B. DuBois's quote at the beginning of this section. In the early 20th century, Binet's assessment techniques were transported to the United States, where they fueled an Eugenics movement that tracked minoritized students to less challenging educational experiences, and privileged white students to more challenging and rewarding experiences. This practice continues to this day, as evidenced by the recent lawsuit in New York City (IntegrateNYC vs. State of NY, 2021). Before describing present day concerns, we will first describe the work of those who sounded the alarm long ago, beginning with the development of professional guidelines for the testing profession.

Standards for Educational and Psychological Testing

In 1952, the American Psychological Association (APA) released *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA, 1952). This document was the first attempt at promoting professional guidelines on test development, use, and evaluation. Two other organizations joined with APA to transform the proposal into the first formal version of standards for the testing industry: the American Educational Research Association (AERA), and the National Council on Measurements Used in Education (which dropped "used" and became NCME in 1961). The first product of this joint effort was *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954). At the time of this writing (July 2025), there have been six versions of these *Technical Recommendations*, the most recent of which are the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; see also AERA et al., 2018 for the Spanish-language version of the

Standards). Despite their efforts to promote more socially just uses of tests, none of these versions can be cited as a helpful treatise on social justice in educational assessment. Nevertheless, we include this work because a review of the evolution of the content of these *Standards* over the past seven decades illustrates increasing concern for the effects of tests on groups of test takers and society.

In the first version of what we today call the *Standards* (APA et al., 1954), it is difficult to find content related to social justice. The words "fairness," "race," and "adverse impact" do not even appear in the document, and "bias" only appears in relation to sampling and statistics. However, these early *Standards* did take a stand against Eugenics by pointing out that using intelligence test to infer innate ability was a misconception. As they put it,

> *another common misconception is that intelligence tests are measures of inherent native ability alone. It would be desirable for manuals of such tests to caution against this interpretation.* (APA et al., 1954, p. 11)

Today, this acknowledgement of this "misconception" seems vacuous, or at best tepid. However, two versions later, APA, AERA, and NCME (1974) explicitly addressed issues of unfairness in test use, and the acknowledgement of such issues required revision of the *Standards.* In describing reasons for the 1974 revision, they claimed,

> *part of the stimulus for revision is an awakened concern about problems like an invasion of privacy or discrimination against members of groups such as minorities or women. Serious misuses of tests include, for example, labeling Spanish-speaking children as mentally retarded on the basis of scores on tests standardized on "a representative sample of American children," or using a test with a major loading on verbal comprehension without appropriate validation in an attempt to screen out large numbers of Blacks from manipulative jobs requiring minimal verbal communication.* (APA, 1974, p. 1)

Thus, by the mid-1970s, issues of social justice in assessment were formally being confronted by the assessment community. With respect to subsequent versions of the *Standards*, the next version (APA, AERA, & NCME, 1985) included two chapters that addressed concerns for "linguistic minorities" and "people who have handicapping conditions," which evolved into a separate chapter on "Fairness

in Testing and Test Use" in the subsequent version (AERA, APA, & NCME, 1999). The 1999 version couched social justice concerns under the rubric of fairness and acknowledged the complexity of "fairness" by stating,

> *concern for fairness in testing is pervasive, and the treatment according to the topic here cannot do justice to the complex issues involved…The Standards cannot hope to deal adequately with all these broad issues, some of which have occasioned sharp disagreement among specialists and other thoughtful observers.* (p. 73)

The current version of the *Standards* (AERA et al., 2014) retained the "fairness in testing" chapter, and described four perspectives of fairness: (a) in treatment during the testing process, (b) as lack of measurement bias, (c) in access to the construct(s) measured, and (d) as validity of individual test score interpretations for the intended uses. Perhaps the biggest fairness issue related to social justice addressed by both the 1999 and 2014 versions of the *Standards* was adverse impact, meaning differential outcomes for subgroups of students based on test scores (e.g., when tests are used for selection into jobs, schools, or competitive programs). In most situations where tests are used for "high-stakes" purposes, African American, Hispanic/Latino, Native American, and other historically minoritized groups have substantially lower acceptance rates, which has led to truncated participation of individuals from these groups in the rewards associated with higher test scores. For this reason, the use of test scores to make such awards has been criticized as supporting systemic racism in education (Hobson, Szostek, & Griffin, 2021; IntegrateNYC vs. State of NY, 2021; Sireci, 2021).

> The AERA et al. (2014) *Standards* addressed this issue by claiming,
>
> *…the* Standards' *measurement perspective explicitly excludes one common view of fairness in public discourse: fairness as the equality of testing outcomes for relevant test-taker subgroups. Certainly, most testing professionals agree that group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias…However, group differences in outcomes do not in themselves indicate that a testing application is biased or unfair.* (p. 54)

The logic underlying this view is, observed mean test score differences across groups defined by racial or other demographic variables may reflect true

differences across groups, rather than imperfections of the measurement properties of the test. However, Helms (2006) pointed out when test scores reflect such differences without explaining the reasons for them in a construct-relevant manner, evidence of *invalidity* persists, and such evidence argues against test use. In a subsequent section, we provide a recommendation for AERA, APA, and NCME for revising this position to emphasize the pursuit of more socially just measurement practices. In our view, the preceding excerpt from the *Standards* sidesteps an important social justice issue that deserves more discourse in the assessment community.

In addition to issues of fairness, the concept of test validity also overlaps with issues of social justice in assessment. As the evolution of the *Standards* indicates, validity evidence can focus on data related to issues of social justice. In fact, in the last two versions of the *Standards* (AERA et al. 1999; 2014), an explicit source of validity evidence based on testing consequences was introduced to expand the conceptualization of validity and validation. Including the consequences of testing as an essential source of validity evidence opened the door to further consideration of social justice issues in assessment.

Sources of validity evidence

It is important to note that the concept of validity and test validation evolved over the six versions of the *Standards*, and the past two versions (AERA et al., 1999; 2014) specified five sources of validity evidence that can be used to evaluate the use of a test for a particular purpose. Three of these sources of validity evidence were discussed in some form since the first version (APA, 1954): validity evidence based on test content, internal structure, and relations to other variables. A fourth source of validity evidence, based on response processes, refers to confirming the cognitive processes intended to be measured are in fact being invoked by the test items. It is the fifth source, however, *validity evidence based on testing consequences*, that overlaps with social justice issues. Ideally, this source of validity evidence should require testing agencies and other test users to provide evidence that the intended consequences of a testing program are being realized, and that (un)intended negative consequences, to individuals or to society, based on the use of test scores are identified and eliminated, or at least minimized. However, the *Standards* fall short of that ideal by requiring only that construct-irrelevant sources of bias be investigated as threats to fairness.

Although concerns over the consequences of testing have been codified into the *Standards* for over 22 years, some researchers have rejected the requirement that consideration of consequences is part of a reasonable validation effort or pertains to validity at all (See Koljatic et al. 2021a, 2021b; Sireci, 2016a, 2016b for discussions). Such a position is at odds with incorporating a social justice perspective into educational assessment, which is why we remain grateful the AERA et al. (1999, 2014) *Standards* rejected such dismissiveness. However, even given the discussions of testing consequences and fairness, the *Standards* remain incomplete with respect to evaluating and providing guidance on social justice issues in educational assessment. Several prominent measurement specialists paid closer attention to these issues and encouraged the profession to do so. It is to these scholars we turn next.

Pioneers of Social Justice in Educational Assessment

We did not complete an exhaustive search of the educational testing literature to identify all discussions of social justice in assessment; however, as students of validity theory and as practitioners who conduct validity studies, several theorists stand out as pioneers in this area. These pioneers include, but are not limited to, Robert Ebel, Samuel Messick, Sylvia Johnson, and Janet Helms. In this section, we provide brief descriptions of some of their writings in this area.

Robert Ebel

In 1963, the Educational Testing Service hosted an *Invitational Conference on Testing Problems* with the closing keynote featuring Robert Ebel, who titled his address "The Social Consequences of Educational Testing" (Ebel, 1963). In this address, Ebel pointed out the testing community was being criticized for "having shown lack of proper concern for the social consequences of our educational testing," and he argued "what testing needs most is a large program of research on its social consequences" (p.131). We agree with his suggestion from 58 years ago and point out it has yet to be acted upon!

Ebel (1963) specified four harmful potential consequences of educational tests: (a) predetermining the status of children based on how they are labeled by test performance, (b) supporting a narrow definition of ability that will "reduce the diversity of talent available to society," (c) placing too much control over education in the hands of the testing industry, and (d) promoting "mechanistic decision

making" (pp. 132–133). Similar to APA et al. (1954), he forcefully argued against using test scores as measures of innate intelligence and claimed "One of the important things test specialists can do to improve the social consequences of educational testing is to discredit the popular conception of the [intelligence quotient]" (p. 135). He also encouraged the use of tests for positive consequences for individuals and society by suggesting, "We should judge the value of the tests we use not in terms of how accurately they enable us to *predict* later achievement, but rather in terms of how much help they give us to *increase* achievement by motivating the efforts of students and teachers" (p. 136). Clearly, Ebel was a precursor to many of the same arguments being made today that educational tests should do more to educate our children (Gordon, 2020; Sireci, 2021).

Ebel (1963) argued if we ignored the criticisms of educational tests, the costs to society would be dire. As he described, "If we ignore them and undertake to manage the lives of others so that those others will qualify as worthy citizens in our own particular vision of utopia, we do justify the concern that one harmful social consequence of educational testing may be mechanistic decision making and the loss of essential human freedoms" (p. 141). In other words, requiring students to answer assessment items in a particular way may overlook the creativity, individuality, and consciousness students possess and bring to the assessment. Ebel's point is that by continuing testing practices in a mechanistic, unchecked manner, we may foster mechanistic education in a way that stifles the diversity of ideas across educators and students. Recognizing that ideas regarding educational constructs differ across race and culture (Malda, van der Vijver, & Tamane, 2010; Randall, 2021), Ebel's call for addressing the criticisms of testing was a call for expanding the worldview of the insular testing profession. For this reason, we consider him one of the pioneers of social justice concerns in educational assessment, and his message endures in contemporary writings in this area (e.g., Dixon-Roman, 2020).

Samuel Messick

In his Presidential Address to the quantitative psychology division of APA, Messick (1975) invited measurement professionals to consider the meaning and values in educational measurement. In addition to focusing on the "construct" measured by a test, he also focused on consequences, pointing out "the social consequences of test use should be weighed against the social consequences of *not* testing"

(p. 962). In his seminal chapter on validity theory (Messick, 1989), he further elevated the importance of considering the consequences of testing by specifying two "interconnected facets" of validity that comprised the "consequential basis" of test validity (p. 20). The first facet he termed the *consequential basis of test interpretation*, which he defined as the "appraisal of value implications of construct label, theory underlying test interpretation, and ideologies in which theory is embedded" (p. 20). The second facet he called the *consequential basis of test use*, which he described as the "appraisal of both potential and actual social consequences of applied testing" (p. 20).

In calling attention to the consequences associated with how test scores are both interpreted and used, Messick underscored how value systems underlie the determination of what is tested, the meaning ascribed to test scores, and how scores are used. His requirement that testing consequences be fully examined was comprehensive, and he provided several examples of how negative consequences or unvalidated test use could lead to injustice at both individual and societal levels. These examples included value-laden labels attached to test scores, adverse impact, and effects of tests on instruction.

Validity evidence in support of test use was not to be taken as an excuse to avoid a comprehensive study of the potential consequences of tests. As Messick (1989) put it, "Even if adverse testing consequences derive from valid test interpretation and use, the appraisal of the functional worth of the testing in pursuit of the intended ends should take into account all of the ends, both intended and unintended, … including… individual, institutional, societal, and systemic effects" (p. 85). In short, Messick's introduction of the consequential basis of test validity compelled the measurement community to discuss whether testing programs foster (or reduce) societal inequities. In essence, he, like Ebel, brought conversations of social justice to psychometrics, albeit indirectly.

Messick's consequential basis of test validity sparked much debate in the testing literature, including two special issues in the NCME flagship journal *Educational Measurement: Issues in Practice* (EM: IP) in 1997 and 1998. The authors in those special issues generally agreed on the importance of evaluating testing consequences, but disagreed whether such evaluation should be considered part of the concept of validity. Nevertheless, as the AERA et al. (1999) *Standards* illustrated (and AERA et al., 2014, too), Messick's call to evaluate consequences—

both intended and unintended—became embodied in best practice guidelines for educational assessment. However, promoting guidelines and standards does not guarantee they will be followed (Johnson, Trantham, & Usher-Tate, 2019). Thus, enforcement of the *Standards* remains a problem (Gitomer et al., 2021).

Sylvia Johnson

Johnson (2000) epitomized the argument that testing consequences are social justice issues in educational assessment. She cautioned against the unintended negative consequences of test-based reform efforts on marginalized students, particularly the use of tests that result in students receiving less rigorous instruction. She claimed such test-based practices had not only negative effects on the individual children, but on society as well because using test scores to make such decisions may result in "a serious loss to society…through failure to identify and develop the real talents of all its members" (p. 151).

Johnson pointed out test-based education reform efforts may lose sight of what is important in education (supporting children to reach their potential) in pursuit of institutional goals. She stated,

> *tests often are advertised as being designed to assess rigorous curriculum standards, but far more attention is typically given to the match between standards and tests than to the essential prerequisite—that is, the match between high standards and instruction for all students being assessed. As a result, the available test products serve mainly institutional needs, offering little benefit to test-takers, teachers, or even to schools in terms of prescriptive information or instructional value.* (p. 155)

She also noted test-based reform efforts were not serving students well due to the harsh and negative messages and actions being sent by policymakers. Her description of this concern falls squarely in the realm of social justice in educational assessment:

> *The language of high standards and testing is often conveyed to the recipients of today's testing products, usually students and their teachers, in a punitive, blame-filled, and even threatening rhetoric which asserts that both have left undone what should have been done and have done what they should not have done. This language further asserts that both groups will reap dire consequences*

> *if test scores do not evidence achievement of high standards. Researchers have shown what tends to happen in many urban schools serving predominantly Black and Hispanic communities under such conditions: Teachers teach to the test in routinized style and emphasize lower order skills in hopes of getting some level of minimal test performance out of their students at the expense of more motivational and interest-eliciting activities that might engage students in a constructive process of learning but which are also risky, may be noisy, take time, and may be more difficult to justify to a supervisor.* (p. 155)

Johnson's description of the effects of tests on historically marginalized students illustrates why studying unintended negative consequences is so important. Studying such consequences is the hallmark of social justice research in educational assessment.

Janet Helms

Like Johnson, Helms (2006) called attention to social justice issues in educational assessment, specifically the issue of adverse impact (the consequences of test score use affecting some groups of test takers more than others). She rejected AERA et al.'s (1999) position that test score differences across groups did not signify bias if they could not be traced to a source of construct-irrelevant variance. In her view, "When test scores that differ by racial groups are used for assessment purposes, resulting decisions regarding members of the lower scoring group are potentially unfair" (p. 845). She pointed out that for historically minoritized students, test scores contained construct-irrelevant variance due to factors that today we would describe as effects of systemic racism. She claimed fair tests involve removing "systematic variance attributable to experiences of racial or cultural socialization" from the scores of minoritized test takers.

Helms (2006) argued the traditional models of test fairness that focus on differential predictive validity miss the mark, because "the focus of these models is on the adverse consequences of using potentially unfair test scores rather than the consequences of such scores" and these models "erroneously treat racial groups as meaningful constructs" (p. 848). Instead of studies of differential predictive validity and measurement invariance, she called for "Replacement of racial and ethnic categories with cultural constructs derived from conceptual frameworks [as] a necessary condition for fair assessment" (p. 848).

Ford and Helms (2012) further pointed out the effects of test-based adverse impact on African Americans claiming,

> *such racial-group test-score disparities mean that typically more than half of African Americans are excluded from a variety of academic and vocational experiences and domains in society beginning when tests first enter their lives. Such exclusion contributes to not only low achievement, but also underachievement; and denied opportunities fuel the greatest educational problem—the achievement gap.* (p. 187)

They further described the lack of voice, and the lack of power, African Americans have in the education system as a clear social justice issue. As they put it,

> *unlike their White counterparts, African Americans…exist in and face racially and economically discriminatory contexts and practices that are not aspects of the socialization experiences of White people. Essentially and undeniably, African Americans exert little influence over the structure and content of the tests on which they are expected to perform as if they are White middle-to-upper class, monolingual Americans.*

Of course, test score differences and adverse impact are not the only manifestations of social injustice in education, or in educational assessment. Ford and Helms helped connect these manifestations to the larger picture or injustices in education and how they interact with the traditional educational assessment practices.

Summary of social justice pioneers in educational assessment

We could elaborate further on the calls for socially just assessment policies made by Ebel, Messick, Johnson, and Helms; as well as by other strong voices that focused on addressing adverse consequences (e.g., Lane, 2014; Linn, 1984; Mislevy, 2018; Shepard, 1993). There have also been professional organizations outside of AERA, APA, and NCME that have promoted more just assessment practices (e.g., the International Test Commission's *Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*, 2018). However, our point in reviewing some of the pioneers in this area is to call attention to the fact that social justice issues are not new. In fact, for decades, prominent measurement specialists have warned of the injustices being caused to both individuals and societies due to the use of educational assessments.

We also acknowledge that the previous calls for more attention to issues of fairness, testing consequences, and validity did not represent a comprehensive discussion of social justice issues in educational assessment. A more comprehensive discussion requires considering consequences at both the individual and societal levels. For example, in describing notions of justice at these levels, Rawls (1999) remarked,

> *justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many.* (p. 3)

In the case of educational assessment, this conflict becomes evident when the use of a test denies access to something that can be considered an individual right, such as the pursuit of a college degree. The social consequences of such test use can benefit colleges and some individuals. However, a focus on social justice requires recognition of consequences that cannot be accepted in any scenario (e.g., denial of fundamental rights). Currently, it appears the measurement profession is stuck on delineating and forming consensus on what are unacceptable consequences associated with educational tests. Forming that consensus, if possible, will require formal dialogues on defining social justice in educational assessment. Understanding and studying the consequences of testing is a prerequisite for a serious conversation in this area. We hope the dialogue can occur soon in the academic and other professional spaces in which measurement specialists operate. Otherwise, it will unfold in the courtroom (e.g., IntegrateNYC vs. State of NY, 2021). In the spirit of promoting this dialogue, in the next section, we offer some suggestions on how we can move forward toward more socially just educational assessment practices.

Promoting Social Justice in Educational Assessment: A Call to Action

As mentioned at the opening of this chapter, social justice has been a difficult concept to define in education. It is likely to prove challenging to define within educational assessment, as well. However, in our view, social justice in educational assessment means developing and providing assessments that value all students; embrace cultural diversity; allow students to access their various funds of knowledge when interacting with the assessment; provide proper support systems for students to successfully access and navigate the assessment; and support the

positive development of students' self-esteem, and their acquisition of academic, occupational, and other goals. These actions acknowledge and honor test taker rights (e.g., APA, 2020) and increase assessment transparency by helping test-takers and educators better grasp and trust the assessment process. Although other measurement specialists may add or delete from our definition, we offer this definition as a starting point, or rather a guidepost, for assessment equity (i.e., supporting the principle that assessments should be fair and considerate of all students' backgrounds and opportunities to learn).

In our previous review of pioneers in social justice in educational assessment, we highlighted some of the social justice issues to be addressed. In this section, we turn to what we can do to become more socially just measurement researchers and practitioners. We believe a social justice perspective in educational assessment must begin with the best interests of each student in mind. These interests include the student's personal and communal values, and so the interests of the communities in which students live must also be front-of-mind. Consideration of these interests involves five key components: equity, access, participation, rights, and diversity. In the next section, we briefly describe actions we in the measurement community can take with respect to each of these components of social justice in educational assessment.

Equity

Like many terms in education, "equity" can mean different things to different people, and is often confounded with "equality." Gordon (1995) distinguished between the two by stating,

> *care must be taken to make clear the difference between equity and equality. Equity speaks to and references fairness and social justice; it requires that the distribution of social resources be sufficient to the condition that is being treated. Equality, on the other hand, connotes sameness and the absence of discrimination.* (p. 363)

Gordon points out equality is not always just, if individuals have different needs, but are given the same resources.

One suggestion we have for making assessments more socially just with respect to equity begins with a request for AERA, APA, and NCME to revise their guidance

on observations of group differences in test scores and adverse impact. The current guidance gives the appearance of "don't shoot the messenger," which is a theme with which many can identify. However, the messenger can also return new information to the original sender, and for far too long, the message returned—that the adverse impact observed based on test scores is unacceptable—has not been adequately responded to. Our suggestion to these organizations is to revise statements in the *Standards* to promote not only the acceptance of responsibility but also to encourage the psychometric community to take collaborative action. Thus, our recommendation is to change a statement such as the following,

> …the *Standards'* measurement perspective explicitly excludes one common view of fairness in public discourse: fairness as the equality of testing outcomes for relevant test-taker subgroups. Certainly, most testing professionals agree that group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias…However, group differences in outcomes do not in themselves indicate that a testing application is biased or unfair. (AERA et al., 2014, p. 54),

to,

> …the *Standards'* measurement perspective acknowledges equality of testing outcomes for subgroups may not always be realized. Thus, when unequal outcomes occur, they should trigger heightened scrutiny for possible sources of test bias, and any discovered sources of bias should be eliminated. Moreover, testing agencies should go beyond the tests themselves and work with educators, employers, and other stakeholders to promote equitable systems of education and employment that do not differ in quality or outcomes for student groups defined by race, ethnicity, culture, and other sociopolitical or sociological characteristics.

This change will acknowledge the point that group differences are not directly caused by tests, which was the point of the wording in the current version of the *Standards*, but it simultaneously acknowledges (a) the problem is not acceptable, and (b) there is an expectation that we help fix it. Our suggestion to add these acknowledgements is consistent with Gordon (1995) who argued,

> *it can be argued that the problems of equity in educational assessment are largely secondary to the failure to achieve equity through educational treatments. However, the fact that these problems of equitable educational assessment are only secondarily problems for assessment does not mean that they should not be engaged by the assessment community, even if they cannot be solved through assessment alone.* (p. 360)

In our view, the AERA, APA, and NCME *Standards* provide an opportunity to describe what testing professionals can do to promote social justice in education. Thus far, the *Standards* have not fully embraced that opportunity. Requiring that testing consequences be studied, without consequences for the testing agencies who ignore them, is a blatant disregard of social justice in educational assessment.

Access and Participation:

Our suggested recommendations to promote greater access and participation in educational assessments are based on the concept of *UNDERSTANDardization* (Sireci, 2020), which extends the concept of promoting access to individuals with disabilities to educational assessments, to promoting access based on all types of test-taker characteristics. UNDERSTANDardization involves keeping the fundamental principles of standardization (i.e., keep everything the same for everyone), but "loosens" what is required to be the same. As Sireci described, testing professionals "must understand the numerous dimensions of heterogeneity that exist within the populations of people we test, and embed that understanding in our standardization processes" (p. 101). Essentially, the goal of this more flexible approach to test development, administration, and scoring is to understand (a) what each student brings to the testing situation in addition to the proficiency measured, (b) how these personal characteristics may interact with testing conditions, and (c) how testing conditions can be sufficiently flexible to accommodate and account for these potential interactions. This understanding will lead to more valid assessment of each individual student's proficiencies.

UNDERSTANDardization uses the principles of standardization to create a more socially just assessment environment by allowing for flexibility in the content and tasks presented to students (i.e., allowing student choice with respect to assessment content and/or context), how they respond (e.g., code-switching between languages, responding orally rather than by writing), and what is

scored (English grammar versus grammar that acknowledges linguistically diverse cultures). By incorporating the principles of UNDERSTANDardization into assessment, students are more likely to "see themselves" in the assessment and feel empowered by the choices they are given.

Participation and Diversity

An important part of UNDERSTANDardization is doing the research to understand the diversity of the population of test takers. This understanding is also key to the larger concept of an antiracist framework for assessment, which Randall (2021) described as an approach to assessment that,

> *requires an explicit confrontation of racism in our assessment practices; and works to disrupt these systems of oppression…An antiracist framework for assessment critically questions the structures and assumptions that make up the judgment of all assessment developers. It is explicit about its politics and its intent to reconstruct hierarchical racial power arrangements that have been historically (re)produced via assessments.* (p. 1)

Thus, the antiracist approach has diversity at its core from the earliest stages of test development. Following these practices will lead to culturally sustaining assessments that de-center whiteness and explicitly acknowledge the beauty and relevance of all cultures, particularly those that have been historically minoritized. A key element of this approach is, rather than screening out test material specific to a particular culture, include it to value that culture. As Randall (2021) argued,

> *the problem is by removing race, or pretending that it does not exist/matter, one is not removing racism. Indeed, such a practice only perpetuates racism as it is simply a proxy for elevating whiteness.* (p. 4)

Another suggestion we have related to diversity is increasing the numbers of African American, Hispanic/Latino, Indigenous, and people from other minoritized groups into the measurement profession. Randall, Rios, and Jung (2021) reported that of 3,124 degrees in the measurement field conferred from 1997 to 2016, only 6.8% were Black and 3.7% were Hispanic/Latino. Clearly, more needs to be done, and the collaboration among the Chan-Zuckerberg Initiative, the Center for Educational Assessment at the University of Massachusetts Amherst, and NCME from 2020–2023 was one important step to increase these numbers. Through this

collaboration, two cohorts of 30 Black and Brown graduate students in educational measurement were awarded travel and mentoring scholarships to attend NCME's annual conference and participate in a program where they were mentored by a Black or Brown Ph.D. scholar. Our suggestion is to improve these efforts and extend them to funding stipends and tuition for Black and Brown students interested in pursuing masters and doctorates in educational measurement.

Call to Action Summary

We have made four modest recommendations for specific actions we can take to promote more socially just assessment practices. These actions are,

(a) Revise the AERA et al. (2014) *Standards'* text related to group differences and adverse impact to encourage testing agencies to work with other stakeholders to reduce inequities and comprehensively investigate validity evidence based on testing consequences;

(b) use the principles of UNDERSTANDardization to provide more flexible test administration practices;

(c) use an antiracist framework for construct definition and test development;

(d) initiate serious recruitment and support efforts to bring Black/Brown colleagues into the measurement field.

It is important for us to add one more action to this list, because this action reflects the good work that has already been done in pursuit of fairness and justice in educational assessment:

(e) continue and expand research on test fairness.

This last action is important because we must acknowledge that the work done in our field for decades to evaluate item and test bias and make tests more accessible is crucial for accomplishing the goals of socially just educational assessment. We must continue such research, but we should not limit ourselves to what has become routine. We must continue to push the boundaries of what we consider to be bias, fairness, and equity; to listen to what the communities within our tested populations consider as fair in *their* contexts, and develop the statistical machinery to evaluate fairness in a situated way. Only then can we claim psychometrics

has done what it can to help all individuals in society attain their academic, occupational, and other personal goals. Furthermore, these actions will lead to assessments of higher quality that support more valid interpretations of students' proficiencies.

Our list of five actions we can take is a brief list, and it is certainly not exhaustive of all that can be done. However, it represents a reasonable starting point for others to join in. We note others are joining in (e.g., Lyons, Johnson, & Hinds, 2021), and we applaud those efforts.

Concluding Remarks

In an earlier section of this chapter, we focused on the work done by APA, AERA, and NCME in promoting guidelines for the testing industry that promote sound and fair testing practices. We also pointed out discussions of social justice and fairness were hard to find in the first two versions of the *Standards* and we recommend improvements to the latest version, which is eleven years old at the time of this writing. A more recent development is relevant and important for us to mention before closing. On October 29, 2021 APA released a formal apology for its longstanding contributions to systemic racism in the profession. Included in the apology was an acknowledgement of the damage done via educational and psychological tests. This acknowledgement is a first step toward reparations, which are also being considered and enacted by APA (See https://www.apa.org/about/policy/racism-apology). Other institutions, such as NCME, should follow APA's lead and acknowledge the misuse of educational tests in a similar manner (Sireci, 2021).

We opened this chapter with a lyric from Peter Tosh. His lyrics are provocative in that he rejects peace, but his point is, by achieving equal rights and justice, peace is a given. We believe by pursuing social justice in educational assessment, we will do our part to move us forward on the path to peace, on which we all should be traveling. By incorporating a social justice perspective to assessment in the service of learning, we not only move closer to equal rights and justice in assessment; we also facilitate the goal of making our assessment systems more transparent, more equitable, and consequentially more valid, for every learner.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7,* 461–465.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, (2, supplement).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Author.

American Psychological Association (2020). *The rights and responsibilities of test takers.* Author. https://www.apa.org/science/programs/testing/rights.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Bell, L. A. (1997). *Theoretical foundations for social justice education*. In M. Adams, L. Bell, & P. Griffin (Eds.), Teaching for diversity and social justice: A sourcebook (pp. 3–15). Routledge.

Bertrand, M., & Marsh, J. (2021a). How data-driven reform can drive deficit thinking. *Phi Delta Kappan, 102*(8). https://kappanonline.org/how-data-driven-reform-can-drive-deficit-thinking-bertrand-marsh/

Bertrand, M., & Marsh, J. A. (2021b). Opting out of standardized tests: The role of parents and their social networks. *American Journal of Education, 127*(2), 231–259. https://doi.org/10.1086/711828

Boyles, D., Carusi, T., & Attick, D. (2009). Historical and critical interpretations of social justice. In W. Ayers, T. M. Quinn, & D. Stovall (Eds.), *Handbook of social justice in education* (pp. 30–42). Routledge.

Dixon-Román, E. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of blackness in measurement. *Educational Measurement: Issues and Practice, 39*(3), 94–96.

Ebel, R. L. (1963). The social consequences of educational testing. Proceedings of the 1963 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.

Ford, D. Y., & Helms, J. (2012). Overview and Introduction: Testing and assessing African Americans: "Unbiased" tests are still unfair *Journal of Negro Education, 81*, 186–189.

Gitomer, D. H., Martinez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: evidence of reliability and validity in the edTPA. *American Educational Research Journal, 58*, 3–31.

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

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

Guthrie, R. V. (1998). *Even the rat was white: A historical view of psychology* (2nd ed.). Boston, MA: Allyn & Bacon.

Helms, J. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845–859.

Hobson, C. J., Szostek, J., & Griffin, A. (2021). Adverse impact in Black student 6-year college graduation rates. *Research in Higher Education Journal*, *39*, 1–15.

IntegrateNYC vs. State of NY (2021). *Index No. 152743/2021.*

International Test Commission (2018). *Guidelines for the large-scale assessment of linguistically and culturally diverse populations*. https://www.intestcom.org/files/guideline_diverse_populations.pdf

Jean-Marie, G., Normore, A. H., & Brooks, J. S. (2009). Leadership for social justice: Preparing 21st century school leaders for a new social order. *Journal of Research on Leadership Education, 4*(1), 1–31.

Johnson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practice, 38*(3), 6–19.

Johnson, S. T. (2000). The live creature and its expectations for the future. *Journal of Negro Education, 69,* 150–158.

Koljatic, M., Silva, M., & Sireci, S. G. (2021a). College admission tests and social responsibility. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12425

Koljatic, M., Silva, M., & Sireci, S. G. (2021b). College admission tests and social responsibility: A response to comments. *Educational Measurement: Issues and Practice*.

Koljatic, M., & Silva, M. (2021). College entrance exams: International perspectives. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana, 58*(1), 1–19. https://doi.org/10.7764/PEL.58.1.2021.2

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33–47.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). *A Call to Action: Confronting inequity in assessment*. https://www.lyonsassessmentconsulting.com/assets/files/Lyons-JohnsonHinds_CalltoAction.pdf.

Malda, M., van de Vijver, F. J. R., & Tamane, Q. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38,* 582–595.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–100). American Council on Education.

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

Porfilio, B. J., Strom, K., & Lupinacci, J. (2019). Getting explicit about social justice in educational doctoral programs in the US: Operationalizing an elusive construct in neoliberal times. *Journal of Educational Foundations, 32*, 104–123.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*.

Randall, J., Rios, J. A., & Jung, H. J. (2021). A longitudinal analysis of doctoral graduate supply in the educational measurement field. *Educational Measurement: Issues and Practice*, *40*(1), 59–68. https://doi.org/10.1111/emip.12395

Rawls, J., (1999). *A theory of justice: revised edition*. Cambridge, MA: Belknap Press of Harvard University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Sireci, S. G. (2016a). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy & Practice, 23,* 319–321.

Sireci, S. G. (2016b). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23,* 226–235.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment, *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G. (2021). Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7–16. https://doi.org/10.1111/emip.12415.

Building Culturally and Linguistically Responsive Workplace Assessments for Learning: An Application to Microelectronics and Engineering Education

Maria Elena Oliveri, Kerrie A. Douglas, and Mya Poe

Abstract

This chapter explores how a Culturally and Linguistically Responsive (CLR) approach can enhance Workplace Assessment for Learning (WAfL) in engineering education. WAfL emphasizes formative, authentic assessment embedded in real-world tasks to support both technical and conceptual learning. We argue that integrating CLR assessment principles—particularly co-design and task contextualization—strengthens WAfL by recognizing learners' diverse ways of knowing, cultural experiences, and problem-solving approaches. Through examples such as interpreting well systems and modeling water treatment facilities, we show how contextualized tasks can reveal important cultural and experiential differences that affect how students engage with engineering problems. Rather than penalizing students for unfamiliar reasoning pathways, CLR-informed WAfL encourages instructors to understand the roots of students' thinking and use it as a springboard for learning. We highlight how formative, evidence-centered design enables students to build metacognitive skills and reflect on their own learning processes. The chapter concludes with an application of these ideas in the context of semiconductor engineering, drawing on a case study for illustrative purposes. This example demonstrates how WAfL and CLR assessment can converge to create equitable learning experiences that prepare students for complex, international engineering work environments.

Principles Addressed:

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

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.

Introduction[1]

In today's rapidly evolving, globalized workplace, it is essential that workers not only have theoretical knowledge and technical skills, but also have crucial professional competencies such as teamwork, professional communication, negotiation, problem-solving, and decision-making skills. These skills are necessary for success in contemporary, multicultural workplaces (National Academy of Engineering, 2004). Within this context, skill development requires employees to be able to understand how co-workers' cultural, linguistic, and educational backgrounds shape their perspectives on and approaches to demonstrating their skills (Jesiak, Zhu, Woo, Thompson, & Mazzurco, 2014; Lohmann, Rollins, & Joseph Hoey, 2006). While training and instruction can help employees acquire these skills, often assessment of these skills is limited. As a result, workplace assessments are needed that not only gauge technical abilities, but also evaluate professional skills that are aligned with culturally and linguistically diverse workplaces (Geisinger, 2016; Oliveri, Mislevy, & Elliot, 2020; Oliveri & Wendler, 2020).

1 Acknowledgments: We wish to acknowledge and thank Drs. Tamara Moore and Eric Johnson for authoring the engineering task included in this chapter. Their expertise in curriculum development and engineering task design has significantly shaped the development of our associated assessments. This contribution highlights the value of collaborative task co-design between curriculum and assessment developers to advance learning and create assessments that support diverse educational settings.

In this chapter, we focus on research on Culturally and Linguistically Responsive (CLR) assessments to prepare learners for the challenges of a diverse, complex modern workplace (Oliveri, 2020; Oliveri, Lawless, & Mislevy, 2019; Oliveri & Wendler, 2020). We focus our chapter in the field of Engineering, which requires workers with technical proficiency, robust professional skills, and a global mentality to thrive in a diverse workplace (Fajaryati, Budiyono, & Wiranto, 2020; Oliveri & Markle, 2017; Paretti & McNair, 2008). Thus, our chapter draws on our prior research on assessing engineering competencies with culturally and linguistically diverse (CLD) populations using a sociocognitive perspective (Douglas, Neumann, & Oliveri, 2023; Oliveri et al., 2019).

Building on our prior work, we identify two significant challenges in engineering education assessment: (a) empowering multiple groups of learners to demonstrate their knowledge and skills in varied ways, and (b) addressing the global demand for engineers with advanced understanding capable of collaborating with globally distributed teams. To address these challenges, we identify two specific frameworks designed to assist in the development of assessments of engineering competencies for CLD populations.

The first framework uses a co-design participatory research approach (Roschelle & Penuel, 2006). A co-design approach guides the assessment development process by involving multidisciplinary subject-matter experts such as industry professionals, educators, and students, to collaboratively develop assessments that consider stakeholders' diverse needs and perspectives to enhance the relevance and inclusivity of the assessments. The second framework is the sociocognitive evidence-centered design (SC-ECD) approach (Oliveri et al., 2019). The SC-ECD approach facilitates evidence-based argumentation, considering the social context of assessment, aligning assessments with educational objectives, and guiding decisions on what to measure and how. To illustrate these frameworks, we use a case study focused on semiconductors and engineering education with the intent to inform workplace readiness and take a holistic approach to assessing both professional and technical skills. By demonstrating the inclusion of a co-design framework and the SC-ECD approach, we show how assessment transparency and assessment equity can be achieved in workplace assessments for learning in engineering education.

Literature Review

Engineering Education in a Globalized Workplace

In a globalized workforce, engineers must be able to apply their technological and scientific knowledge in engineering practices, which involve professional skills such as collaborating with distributed, multicultural teams, problem-solving, and design decision-making (Ball et al., 2012; Fajaryati et al., 2020; Jesiak et al., 2014; Lohmann et al., 2006). Globally, engineering educators strive to equip engineering students with the skills to integrate diverse perspectives from various stakeholders and successfully execute projects in multicultural contexts (Kim & Care, 2020; National Research Council, 2012; UNESCO, 2016). For example, Auer and Rüütmann (2021) emphasized the need for engineers to address complex, ill-defined problems, collaborate effectively in multinational teams, and stay ahead of technological advancements. As a result, merely having broad knowledge is insufficient. Engineers must have a deep understanding of their discipline, professional skills, and ability to work with CLD populations as well as the ability to work with unfamiliar contexts beyond textbook problems (ABET Criterion 3 Outcomes 1–7; Jorion et al., 2015). In response to such changes, educational practices have shifted from expecting rote memorization and basic skill application to encouraging students to use their knowledge alongside professional skills to complete complex tasks in the global engineering workplace (Kim & Care, 2020; Pellegrino, 2012).

To achieve these educational shifts, assessment approaches must align with the competencies educators aim to impart and recognize the diverse ways students can demonstrate their abilities in the complex modern engineering workplaces. Our work is meant to augment the research on assessment in engineering education that has been informed by ABET accreditation standards in the U.S. (Olds, Moskal, & Miller, 2005) as well as a variety of approaches that include problem-based learning and authentic learning (Merzdorf et al., 2023; Paretti, 2006) as well as reflection (Cajander, Daniels, McDermott, & von Konsky, 2011). This focus is crucial not only to meet the needs of educators, psychometricians, and professionals committed to fostering inclusive, equitable assessments for CLD populations but to also proactively confront potential biases ingrained in curriculum and assessment development, which might disadvantage marginalized learners and contribute to societal disparities (Lyons, Oliveri, & Poe, 2025).

Culturally and Linguistically Responsive Teaching and Assessment

In recognition that many traditional approaches to teaching did not serve historically minoritized students, CLR research and teaching practices and its variants (e.g., culturally responsive teaching, culturally sustaining pedagogy, and culturally relevant pedagogy) emerged in the 1990s (Gay, 2002, 2013; Ladson-Billings, 1995a, 1995b; Lee, 1998; Qualls, 1998). CLR approaches seek to include students' rich cultural and linguistic backgrounds such that those backgrounds are validated in school contexts. For instance, Ladson-Billings (1995a, 1995b) advanced that CLR assessments connect learning to students' home and community cultures and situate content within meaningful contexts, leveraging students' funds of knowledge. To make assessments more relevant and meaningful, Moll, Amanti, Neff, and Gonzalez (1992) further highlighted the importance of drawing on the concept of "funds of knowledge" to enable students to use their accumulated cultural and linguistic experiences to navigate their social worlds and to develop and use materials that connect with students' home and community experiences.

CLR-informed approaches to teaching include a range of practices, including allowing students to draw on their own background experiences and knowledge in classroom contexts, acknowledging and valuing the range of linguistic practices students bring to school systems, and developing students' critical consciousness to solve real-world problems, especially those related to social inequities. In this way, cultural and linguistic variation were central to educational practice and recognized that students could draw on that variation in developing and critiquing knowledge within the context of high-achieving classroom learning.

For our purposes, we see promise in two hallmarks of CLR-informed assessment. First, CLR assessments do not merely seek to represent learners' identities in test items, but rather include co-design in assessment development. In co-design, members from different cultural groups are included in the assessment-design team. This step requires being mindful of power dynamics across team members to evenly represent different cultural groups. Hood (1998) also suggested diversifying expert groups to co-define assessment constructs, minimize rater bias, and pilot tasks to evaluate their psychometric properties. Qualls (1998) underscored the crucial role of co-design and collaboration across all assessment development stages within evenly distributed cultural groups in an assessment design team. More recently, Randall (2021) emphasized the need for an anti-racist assessment approach to deliberately include Black, Indigenous, and People of

Color's sociocultural identities throughout the assessment process, from the planning to the development phases.

Second, CLR assessment values task contextualization that integrates learners' ways of knowing and learning (Solano-Flores & Nelson-Barber, 2001). As noted by the National Academies of Sciences, Engineering, and Medicine (2018), student learning is deeply influenced by social and cultural contexts. These factors need to be considered in assessment design and score interpretation, rather than reflecting only the values of test designers (Nasir & Hand, 2006). As a result, Hood (1998) suggested considering students' linguistic and cultural backgrounds for assessment validity, advocating for the creation of culturally specific tasks grounded in content validity.

Bennett (2023) and Randall et al. (2022) suggested that contextualizing academic knowledge and skills within students' lived experiences and frames of reference enhances personal relevance, increases engagement, and better measures what students know and can do. Without considering linguistic diversity, assessment passages and questions can become a source of construct-irrelevant variance, becoming less accessible, leading to score misinterpretation, disconnection, and increased cognitive load for CLD learners (O'Dwyer, Sparks, & Nabors Oláh, 2023; Oliveri, 2019; Oliveri et al., 2019). A CLR assessment approach aims to recognize and incorporate test takers' diverse cultural backgrounds and linguistic patterns when responding to test items or constructed-responses, rather than treating them as errors (Mislevy, Oliveri, Slomp, Wolf, & Elliot, 2025).

The implications of CLR assessment extend beyond test design to include score interpretation. Evans (2021) and Solano-Flores and Nelson-Barber (2001) argued that CLR assessment practices should inform how scores are interpreted and used. For example, most methods used to analyze differential item functioning (DIF) often assume cultural neutrality. These methods identify items that perform differently across groups, but do not address the cultural assumptions embedded in the items themselves. When cultural assumptions are identified within test items, such as an assumption that test takers are familiar with municipal water supplies, we can begin to conduct more nuanced analyses of differences in test scores.

For instance, to continue our description of the water example: Consider a test item that describes the processing of municipal water supplies. For rural students accustomed to well-water systems, this context may require additional cognitive effort to imagine how municipal water is stored and treated. The complexity increases if the item asks a policy-related question about EPA standards for contaminants. In many rural areas, well water is not subject to the same regulatory requirements, the cost of testing and treating well water typically falls on the well owner. In contrast, municipal water supplies are regularly tested and the cost of water treatment is often part of the local tax burden. What, then, can we infer about the framing of a test item on water treatment using a municipal water context? Such framing may inadvertently introduce construct-irrelevant variance for certain test-taker groups during test design—for example, should we delay water testing because we cannot afford to treat the water if there are contaminants? Score differences arising from these embedded cultural assumptions could go unnoticed in traditional analyses that focus on disaggregation by gender, race, ethnicity, or other commonly used demographic characteristics. CLR assessment practices encourage deeper exploration of such cultural contexts to enhance fair and valid score-based inferences. By explicitly considering such differences, developers can create varied tasks that cater to the specific needs of CLD populations and develop more nuanced ways of analyzing test scores.

CLR Assessment Frameworks for Workplace Assessments for Learning

In the context of the workplace, workplace assessments for learning (WAfL) can be designed to support learning of complex constructs like teamwork, communication, metacognition, problem-solving, and decision-making (See Wiliam, 2011, for a discussion of AfL). Prior research on WAfLs highlight their importance across occupations, particularly in science, technology, engineering, and math (STEM; Douglas et al., 2023; Oliveri et al., 2021). These assessments rely on next-generation item types such as simulations, scenario-based assessments (SBAs), and situational judgment tests (SJTs) to provide authentic learning experiences immersed in realistic workplace settings (Merzdorf et al., 2023). WAfLs are designed to allow test takers to collaborate, innovate, and apply their knowledge to solve real-world problems, preparing them for advanced technical workplaces (Douglas et al., 2023).

The hallmark of WAfL, like other forms of AfL, is that they are designed to support learning, that is, they "provide information about what kinds of instructional activities are likely to result in improving performance" and "the learner engages in actions to improve learning; this may be undertaking the remedial activities provided by the teacher, asking a peer for specific help, or reflecting on different ways to move her own learning forward—after all, the best designed feedback is useless if it is not acted upon" (Wiliam, 2011, p. 12). In engineering contexts,

> *assessment feedback guides engineering students to monitor and regulate their learning. Lizzio and Wilson (2008) state that the implication of feedback on students' learning is in the identification of their strengths and weaknesses in their performance. Assessment feedback in this sense can develop a kind of self-regulation among students for the improvement of their learning (Nicol, 2009).* (Subheesh & Sethy, 2020, p. 12)

If we return to the water treatment example, we can see the potential of a WAfL approach. Rather than waiting for summative test results to identify conceptual issues, an engineering professor might use a workplace simulation to assess computational accuracy and conceptual understanding. For example, in a water treatment facility simulation, students could be asked to calculate flow measurements and demonstrate their conceptual understanding of how water treatment facilities function, including the role of external government agencies in standard-setting. The simulation becomes an even more powerful illustration of WAfL if students assess their own learning beforehand, developing metacognitive abilities to distinguish between computational errors versus conceptual misunderstandings.

From a CLR perspective, students' background experiences with wells may shape their thinking in different ways—some may focus on depth, material composition, or flow rate, based on local use or community knowledge, rather than thinking about wells in terms of regional water supply systems or contaminants that must be monitored. These variations highlight how learners' prior knowledge and cultural context influence the framing of technical problems. When embedded in WAfL tasks, such examples allow instructors to surface and address these differences productively, supporting equity in assessment and instruction.

A WAfL approach ensures that the task is not only technically rigorous, but also relevant to real-world scenarios engineers may face—particularly in complex, international, and interdisciplinary settings. In addition to thoughtful task design, WAfL requires the generation of scores through authentic, open-ended tasks. These tasks reflect real-world problems that lack clear-cut answers and can be approached from multiple directions. Gutiérrez Ortiz, Fitzpatrick, and Byrne (2021) note that such tasks often present either insufficient information—prompting students to determine what is needed and how to obtain it—or too much information—requiring students to sift through data and identify what is relevant. These conditions support construct representation and encourage diverse problem-solving strategies. To further illuminate student thinking, instructors may prompt learners to describe their initial steps in problem solving. Kalyuga and Sweller (2004) found that novices tend to rely on trial-and-error or rigid procedures, while experts employ more strategic, high-level approaches. These think-aloud methods can reliably differentiate between varying levels of competence and provide instructors with formative insights into student understanding—one of the key goals of WAfL. As we illustrate later, these considerations are important to acknowledge from a CLR perspective to ensure that students' diverse ways of thinking and problem-solving are appropriately recognized and not misinterpreted as errors.

In sum, we see much potential in applying a CLR assessment approach to WAfL by advancing two hallmarks of CLR assessment: co-design and task contextualization. These principles support the integration of learners' varied ways of knowing and learning within an evidence-centered framework. In the next section, we illustrate the application of the ideas within the context of a semiconductor engineering program.

Case Study: An Application of the CLR Assessment Frameworks to Semiconductor Engineering

To address the growing challenges of rebuilding a secure domestic microelectronics industry, the U.S. and its allies require a skilled workforce with advanced multi-dimensional skills. Essential to this endeavor is the development of instructional materials and assessments that bridge the gap between workplace and academic cultures, to promote workforce development initiatives, prepare recent graduates for work, and stimulate interest in industrial careers. Addressing this goal necessitates a strategic CLR WAfL assessment development approach that includes understanding the unique barriers underrepresented students face and designing assessments that minimize cultural and linguistic validity issues.

Framework 1: A Co-Design Approach

Roschelle and Penuel (2006) described co-design as a collaborative process involving stakeholders such as teachers, employers, researchers, and developers working together in defined roles to create educational innovations, prototypes, and assessments tailored to specific workplace or academic needs. Unlike traditional top-down approaches typically used in large-scale assessment design where the psychometrician designs tests, the co-design process actively engages stakeholders, empowering them with ownership and agency over resulting products.

A co-design approach can help create more culturally responsive and equitable assessments. By involving members of a population who are often marginalized, a co-design framework can better address the power dynamics and biases inherent in traditional assessment practices. It allows for the development of assessments that recognize and value the diverse cultural and social contexts in which learning occurs, leading to fairer and more accurate measurement of knowledge and skills. Suárez-Álvarez, Oliveri, Zenisky, and Sireci (2024) demonstrated how co-design can facilitate early identification of potential pitfalls and biases in assessment design. By involving industry specialists, academic institutions, and adult learners, they identified assessment needs within the Adult Skills Assessment Program through focus groups and literature reviews. This process helps determine assessment priorities, contextualize tasks within real-world scenarios, and develop materials that reflect diverse learners' lived experiences.

Table 1 outlines the key stakeholders suggested for co-designing assessments in engineering education programs focused on semiconductors. The table identifies each stakeholder, describes their role in the assessment-design process, and highlights their importance in ensuring that assessments are CLR and inclusive.

Table 1.
Co-Design Stakeholders and CLR Assessment Considerations

Stakeholder	Stakeholder Roles and Relevant CLR Considerations
Academic Institutions, Faculty	Design curriculum & learning objectives with industry, educators, community representatives, and learners; align educational content with industry needs. Include diverse faculty members for inclusivity and curricular relevance to industry.
Federal Employers, Industry Partners, & Workforce Development	Shape assessments to reflect public sector needs with companies ranging in size and culture; reflect needed KSAs and inform real-world scenario development. Include individuals from underrepresented groups. Ensure career resources are inclusive and address diverse career pathways.
Current and Prospective Students	Offer feedback on learning experiences, participate in piloting assessments. Collaborate with peers and instructors, incorporate examples and scenarios from various cultures; advocate for multiple assessment formats (e.g., oral presentations, practical projects and accommodate different learning styles).
EdTech Companies/ Technologists	Develop technology-based tasks and tools for assessments. Use analytics and reporting tools to track the performance of different groups and provide support as / where needed. Develop assessments that use diverse cultural references and are more relatable. Create a more personalized assessment experience. Develop prompts that account for a range of cultural norms and values including analyzing responses for potential cultural biases and adjusting scoring algorithms accordingly.
Cultural and Linguistic Consultants	Advise on strategies to integrate CLR principles into assessment design and implementation, help eliminate biases and ensure equitable assessment practices.

Co-design starts with the goal of creating tangible innovations and includes documenting the classroom context through instructor engagement and accommodating flexibility based on teacher input. Effective collaboration is ensured through strong facilitation and well-defined roles, while central accountability for quality assurance is maintained by the principal investigator or researcher. This collaboration fosters the creation of innovative forms of assessment, such as simulations, SBAs, problem-based learning, and SJTs. These methods evaluate the multi-dimensional skills required in the microelectronics field, providing authentic learning experiences that mirror workplace realities. Faculty from various backgrounds, students with different learning needs, and industry professionals from diverse sectors all contribute to a more comprehensive and inclusive curricular and assessment design process. This diversity enriches the development process and ensures that the resulting assessments are equitable and effective.

Creating partnerships between academia and industry further strengthens the alignment of educational outcomes with workforce requirements. Co-designing assessments with industry partners helps educators understand the skills valued in the workplace and incorporate these into assessment tasks. By integrating these considerations, a semiconductor engineering program can more effectively prepare a diverse and highly skilled workforce and diversify the student population to include more CLD learners.

In such programs, a strategic and CLR co-design approach brings together stakeholders from industry, educators, and students to design assessments for learning to help ensure the training and assessment program meets the demands of the microelectronics industry. Co-designing assessments that are CLR not only enhances the learning experience but also strengthens the program's overall impact because it addresses employers' concerns about the potential of employees to work in global, complex workplaces. In other words, it helps to close the gap between academia and the workplace. However, co-design alone is not sufficient to address all test design considerations. Thus, in the next section, we illustrate our second framework, the CLR expanded ECD (e-ECD) Assessment for Learning approach, which we refer to as the expanded sociocognitive ECD (SC-ECD) approach. It builds on the assessment for learning approach proposed by Arieli-Attali et al. (2019) and includes CLR considerations for constructing assessments that accurately measure the intended skills and knowledge while being sensitive to cultural and linguistic differences. The SC-ECD assessment for learning approach

combines co-design and CLR e-ECD approaches to help ensure the assessments for microelectronics engineering are both inclusive and methodologically rigorous, integrating cultural and linguistic responsiveness with technical rigor.

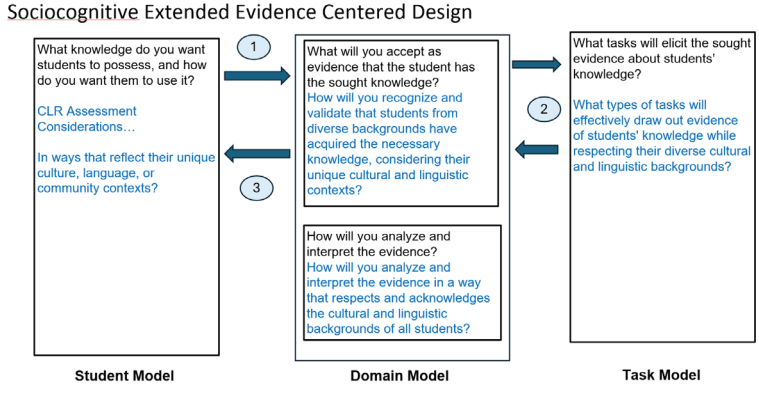
Framework 2: The Expanded Sociocognitive Evidence-Centered Design Framework (SC-ECD)

The e-ECD approach (Arieli-Attali et al., 2019) complements a co-design framework. The e-ECD approach builds on the original Evidence-Centered Design (ECD) framework conceived by Mislevy, Almond, and Lukas (2003). ECD has been applied to various contexts, including engineering assessments and knowledge-in-use tasks that incorporated disciplinary ideas, cross-cutting concepts, and science practice (Harris et al., 2019; Pellegrino et al., 2014). Arieli-Attali et al.'s e-ECD framework extends traditional ECD by integrating learning aspects with assessment elements and considering both cross-sectional and longitudinal perspectives on learning. In the framework referred to as SC-ECD, Oliveri et al. (2019) expanded ECD in alignment with CLR assessment design and development principles.

SC-ECD builds on e-ECD to guide assessment design by addressing evidentiary aspects relevant to the construct and various assessment facets including considerations for learning and assessment, digital instructional content, and measurement models for learning. (Figure 1 shows components of the e-ECD approach; the relevant CLR assessment considerations are shown in blue text).

Figure 1.

ECD Model Adapted for CLR Assessments



Source: Adapted from Pellegrino et al., 2014

Table 2 offers example answers to the questions listed in Figure 1 for each of the components of the e-ECD model along with CLR assessment considerations. In alignment with our goals to design WAfLs for CLD learners, the focal tasks include innovative, workplace-relevant constructs including teamwork, communication, and problem-solving. In the task, we also suggest creating opportunities to engage with fictitious clients in diverse company settings. We elaborate on task design decisions when we illustrate the application of our frameworks to engineering education and semiconductors.

Table 2.

Components and Examples of the CLR Assessment e-ECD Model

Key Component	Design Principle	Elaboration	Cultural/ Linguistic Considerations
Learning and assessment	Assess integrated technical and professional skills to provide learners opportunities to practice skills bridging the academic and workplace contexts	Contextualize tasks in realistic, authentic scenarios that are relevant to modern workplace settings	Engage in varied cultural contexts with peers from different cultural and linguistic backgrounds
Digital instructional content	Create authentic learning scenarios and provide opportunities for hands-on, project-based learning	Include varied forms of communication, negotiation, problem-solving, and teamwork	Include scenarios across different contexts including rural or urban contexts
Measurement models for learning	Cognitive diagnostic models, open-ended responses, engineering projects, or rubrics to provide feedback to learners	Use analysis methods that enable diverse responses that acknowledge varied ways of item responding	Allow for diverse linguistic, cognitive, and substantive patterns to be used in responding to tasks

An Example of CLR Assessment Task Design and Scoring in Engineering Education and the Semiconductor Industry

Table 3 shows a task related to developing a stroke recovery system and designing an Application Specific Integrated Circuit (ASIC) at a Semiconductor Company. To deepen students' understanding, this task prompts students to tackle technical challenges in ASIC design for medical devices, explore stroke rehabilitation principles, navigate regulatory requirements, and engage in industry collaboration. The task is contextualized to enhance engagement by grounding it in real-world applications like improving stroke patient recovery. This practical focus makes engineering concepts tangible and relevant, likely boosting student interest. Students learn skills such as ASIC design, low-power optimization, manufacturing processes, quality control, and meeting regulatory standards—also key for semiconductor careers. It also emphasizes collaboration, project management, and communication, crucial in any workplace. The task assesses technical proficiency, problem-solving, regulatory compliance, and innovation. Students analyze neurostimulation technology in stroke rehabilitation, address technical challenges, and propose solutions by demonstrating creativity and interdisciplinary collaboration. However, the task lacks CLR assessment elements and opportunities for students to engage in CLR learning. Including diverse cultural and linguistic perspectives would enhance its relevance and inclusivity, ensuring all students can relate to and benefit from the learning experience.

Table 3.

Example Assessment for Learning Task for the Semiconductor Industry

Chicago Neuroscience Selects Semiconductor Company to Develop ASIC for New Stroke Therapy Neurostimulation Device; Technology Intended to Improve Patient Recovery

FRANCIS, Indiana—Semiconductor Company, a leading designer and manufacturer of state-of-the-art integrated mixed-signal and structured digital products for the automotive, medical, and industrial markets. Today Chicago Neuroscience, Inc. (CNI), a medical device company, selected Semiconductor Company to design and manufacture the Application Specific Integrated Circuit (ASIC) for the CNI Stroke Recovery System. Hayden Levy, MD, a neurosurgeon who is CNI's Medical Director, commented on CNI's investigational therapy: "After a stroke, the brain attempts to compensate for the damaged area by reorganizing through a process known as neuroplasticity. However, many survivors of stroke remain impaired for the rest of their lives. We have spent years developing a device system which is intended to help improve function in stroke survivors months or even years after their strokes." The American Stroke Association estimates that about 700,000 people in the U.S. experience a stroke every year. The estimated direct and indirect healthcare costs related to stroke was over $56 billion in 2005. Initial research has demonstrated that cortical stimulation of healthy brain tissue near the area damaged by the stroke, in combination with rehabilitation, may facilitate neuroplasticity and improve function. CNI's stroke therapy device involves the precise delivery of low levels of electricity to the surface of the brain (the cortex) via an implanted stimulator system. Results from two feasibility studies presented at medical congresses this year assessed the safety of cortical stimulation in the rehabilitation of chronic stroke patients and suggest a greater improvement in recovery of hand/arm function compared to controls. The company is currently enrolling patients in a larger pivotal clinical study to confirm these results. The ultra-low-power consumption ASIC for the CNI system is being designed and produced by Semiconductor Company at the company's Fort Wayne, Indiana design center and Dallas, Texas manufacturing facility. "After an intense evaluation of a number of potential partners serving this market, it became clear that Semiconductor Company offers outstanding capabilities, custom production processes, and experience as a manufacturer of ASICs for implantable medical devices," said Justine Wyler, Ph.D., president and CEO of Chicago Neuroscience, Inc. "Numerous medical device companies have come to rely on Semiconductor Company to meet the critical technical requirements of implantable devices," said Hayden Barnes, vice president of the medical and wireless product line at Semiconductor Company. "We are very excited to partner with CNI to bring their potentially life-changing product to patients. We have worked diligently to develop a successful track record in delivering critical integrated circuits on time and above the required quality levels. We look forward to continuing that success with CNI."

AfL task-relevant questions include asking students: 1. Write an explanation for why Chicago Neuroscience, Inc. (CNI) selected Semiconductor Company as a partner? 2. Thinking about how engineers design digital circuits and try to optimize the circuit designs: What features of the circuit are you trying to optimize? Why do you try to optimize those features? 3. For each feature optimized, explain why you think this is an important feature to be optimized? 4. When designing a circuit, it is crucial to ensure that your code functions properly. Therefore, compile a list of strategies to debug circuit code when it malfunctions.

Table 4 shows the ways that the task could be extended to include CLR assessment considerations. That is, while keeping in line with the task shown, this activity could be extended to encourage students to think about whether cultural or linguistic differences would emerge if two or more different cultural groups were to engage with this task. The CLR assessment considerations should be developed through the co-design process in which students and industry stakeholders participate in the task design process.

Table 4.

Revised Task with Cultural and Linguistic Diversity Considerations

Skills/Learning Objectives	Items Relevant to the Skill	Task Design Considerations	CLR Assessment Considerations
Understanding Cultural Perspectives	U.S. Perspective: Focus on the high prevalence of stroke and significant healthcare costs. CLR assessment Perspective: Discuss cultural attitudes toward technology and medicine.	Emphasize innovative medical devices to improve long-term recovery and reduce healthcare costs.	Highlight diverse cultures' advanced healthcare system with a focus on preventative care and rehabilitation. Discuss the acceptance and integration of new medical devices.
Technical Design and Development	U.S.-based Engineers: Design ASIC to meet stringent U.S. regulatory standards and ensure compatibility with existing infrastructure. CLR assessment/ CLD Engineers: Adapt ASIC design to be compatible with other cultures' healthcare practices.	Focus on energy efficiency and integration with other medical technologies used in American hospitals.	Consider local regulatory requirements and preferences for medical device design. Possibly integrate traditional medicine approaches.

Table 4. (continued)

Skills/Learning Objectives	Items Relevant to the Skill	Task Design Considerations	CLR Assessment Considerations
Collaboration and Communication	Intercultural Collaboration: Form interdisciplinary teams with members from both the U.S. and other cultures. Meetings and Documentation: Conduct meetings that respect different communication styles.	Share best practices and innovative ideas. Provide clear agendas and summaries in both English and other languages.	Use multilingual support tools to facilitate communication and ensure all team members can contribute effectively. Ensure technical documentation is available in different languages, using culturally relevant examples and terminologies.
Patient-Centric Design	American Patients: Consider expectations and needs of stroke patients, including a preference for advanced technology. CLD Patients: Adapt designs to fit local practices.	Focus on comprehensive aftercare.	Consider the holistic approach preferred by patients from different cultures, which might include complementary therapies alongside high-tech solutions.

Table 4. (continued)

Skills/Learning Objectives	Items Relevant to the Skill	Task Design Considerations	CLR Assessment Considerations
Example Questions:	Technical Challenge: How would you design the ASIC to comply with FDA regulations and integrate with U.S. healthcare technologies?	Focus on meeting FDA's medical device regulations.	Adapt ASIC design to meet other culture's healthcare regulations and consider local medical practices, including the potential integration of traditional medicine.
	Cultural Considerations: Discuss challenges and benefits of introducing advanced neurostimulation devices to American stroke patients.	Address potential challenges in technology adoption.	How might different cultural attitudes towards rehabilitation and technology influence the design and acceptance of the neurostimulation device?
	Collaboration and Communication: Describe strategies for effective communication and collaboration between U.S. and CLD teams.	Ensure cultural and linguistic differences are respected and leveraged.	Enhance the project by leveraging cultural and linguistic differences.

Note. In this scenario, Chicago Neuroscience, Inc. (CNI) has selected Semiconductor Company to design and manufacture the Application Specific Integrated Circuit (ASIC) for their new stroke recovery system, intended to improve patient recovery through neurostimulation. The project involves interdisciplinary collaboration and understanding diverse cultural perspectives on medical practices.

By incorporating CLR considerations into the task design, students can better articulate, understand, and critique the global context of engineering projects and develop skills that are crucial for working in diverse, interdisciplinary teams. They can also learn to develop scoring and evaluation methods of such skills through the rubric-design process.

CLR Scoring Processes

Evaluating responses to open-ended problems presents a challenge from a CLR assessment perspective. A central issue is designing formative learning rubrics that can capture the range of possible responses to the CLR-informed task (Wylie & Lyon, 2016). One effective strategy is to score the complexity of students' solutions relative to all possible solutions. For instance, Fortus et al. (2019) utilized a rubric to evaluate students' three-dimensional learning about energy, focusing on the integration of core notions, scientific and engineering practices, and cross-cutting concepts. Descriptive self-assessments, where students explain and critique their problem-solving process, can also provide insight into their understanding and differentiate between novice and expert strategies. The point here is that a CLR assessment perspective to scoring values process as well as product. It is often more useful in classroom contexts to understand how students came to a solution rather than merely scoring the solution itself. From a CLR assessment perspective, scoring process and product provides greater transparency in assessing student learning and builds equity in the classroom assessment context because students are not merely rewarded for getting a single right answer.

Using a flexible performance-based rubric, instructors, students, and stakeholders can work together to develop rubrics that specifically foreground the issues they identified as CLR values in the task design. Co-design is critical here for three reasons. First, co-design of the rubric in relation to co-designing the task ensures a closer alignment of the scoring to the initial prompt. In the case of the revised task with cultural and linguistic diversity considerations (Table 4), the rubric would likely include facets related to understanding cultural perspectives as well as technical design and development, collaboration and communication, and patient-centric design (Table 5).

Second, because rubrics often use vague terms like "clarity," a column may be added to the rubric in which scorers work together to define, in their own words, the trait that they are scoring. In this way, definitional variations in concepts such as "technical accuracy" come to light. Broad concepts such as "communication" would be defined in linguistically correct ways through the use of terms such as cohesion (connection of sentences to each other), stylistic variation (variation in structure and length of sentences), lexicon (use of technical language), surface error (typographical, capitalization, and punctuation errors) (Gopen & Swan, 1990: Irish, 2015; see also Steiss et al., 2024, for an example related to writing in history). This approach would ensure stereotypes related to language use would be minimized in the design of the rubric. Such stereotypes are most likely to affect multilingual writers when scorers are unable or untrained to distinguish different types of linguistic features in texts. By comparing definitions and working together to develop a shared understanding of a specific trait, students, teachers, and other stakeholders come to better understand that there is often not a shared assumption that informs scoring. Instead, consensus must be built by recognizing different viewpoints.

Third, in a CLR assessment approach, teachers (or graduate assistants) are not the sole arbitrators in the assessment process. Instead, assessment is a team effort where students assess each other through peer review and students self-assess their own work. Teachers and/or industry stakeholders also participate in the process. Invariably, there will be disagreement. Rather than seeing that disagreement as evidence of construct-irrelevant variance or lack of reliability, such disagreement becomes a source of awareness about different viewpoints and how to negotiate those different viewpoints in a complex workplace. This approach allows students to demonstrate their knowledge and skills in various ways, ensuring a fair and comprehensive evaluation. The collection of evidence related to the varied viewpoints of scores becomes, also, a powerful source of evidence related to the assessment process and can help inform future task designs as well as be used in crafting validity arguments related to response processes.

Table 5.

4-Point Rubric for Workplace Communication in a Microelectronics Problem

Criteria	Level 1 (Basic)	Level 2 (Developing)	Level 3 (Proficient)	Level 4 (Advanced)
Technical Design and Development: Circuit optimization and features	identifies features of the design but cannot frame those choices from an engineering perspective	identifies features of the design solely from an engineering perspective	identifies features of the design from an engineering perspective with an awareness of their effects on patients	identifies features of the design from an engineering perspective with an awareness of their effects on patients and considers possible solutions to negative effects on patients
Technical Design and Development: Code debugging	identifies features of the debugging process but cannot frame those choices from an engineering perspective	identifies features of the debugging process from an engineering perspective	identifies features of the debugging process from an engineering perspective with an awareness of their effects on patients	identifies features of the debugging process from an engineering perspective with an awareness of their effects on patients as well as considers possible solutions to negative effects on patients
Technical Accuracy: Lexicon	Uses basic technical terms inaccurately	Uses technical terms correctly most of the time	Uses technical terms accurately and appropriately	Uses technical terms accurately, appropriately, and with confidence

Table 5. (continued)

Criteria	Level 1 (Basic)	Level 2 (Developing)	Level 3 (Proficient)	Level 4 (Advanced)
Organization	information is presented in a way that does not resemble a workplace email, including addressee, subject line, and purpose statement	information is presented in a way that resembles a workplace email, including addressee, subject line, and purpose statement but may be too long or not sufficiently detailed	information is presented in a way that resembles a workplace email, including addressee, subject line, and purpose statement and that provides detailed yet scannable response to the task	information is presented in a way that resembles a workplace email, including addressee, subject line, and purpose statement and that provides detailed yet scannable response to the task. E-mail may include a request for next steps or reply from management
Understanding Cultural Perspectives	Shows little awareness of cultural and linguistic differences	Shows some awareness of cultural and linguistic differences, such as awareness that not all readers may be from a single context	Demonstrates awareness and sensitivity to cultural and linguistic differences, such as awareness that not all readers may be from a single context and strives to use a consistent lexicon and sentence structure to aid readability	Demonstrates a deep understanding and sensitivity to cultural and linguistic differences, effectively incorporating them into communication. For example, the author shows awareness that not all readers may be from a single context and strives to use a consistent lexicon and sentence structure to aid readability. The author suggest ways that the team can accomplish design goals across time zones and other challenges brought on by the demands of the global workplace

Conclusion

In this chapter, we sought to bridge assessment design between the academic and workplace contexts, by offering an exploration of their application to the engineering context. By addressing the limitations inherent in traditional curricula and assessment models, we advocate for innovative WAfL approaches grounded in CLR assessment principles. CLR assessment principles are not merely about acknowledging differences in a global workplace, rather embedding that awareness within the ways that assessment practice is undertaken. Specifically, we argue that co-design and sociocognitive evidence-based models that are informed by CLR assessment principles provide the expanded ways of thinking that are needed to address assessment of learning in contemporary engineering practice and help students bridge academic and workplace contexts. Such practices result in assessment that is not just more equitable for diverse learners but also more transparent as it brings more stakeholders to the table.

In the end, the engineering education context is a robust site for the development of CLR WAfL. Engineering curricular development has long been attuned to workplace realities, which has resulted in the ready adoption of teaching approaches such as project-based learning. Yet, the assessment of learning in engineering education has often lagged behind pedagogical innovation. Moreover, the inclusion of CLR principles in engineering education has been limited, in part, because of the separation of technical skills training and "soft skills" training (e.g., teamwork, communication, metacognition, problem-solving, and decision-making). Using co-design and evidence-based measures, CLR WAfL brings together these two elements for the improvement of learning and assessment design.

References

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology, 10*, 853. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00853/full

Auer, M. E., & Rüütmann, T. (Eds.). (2021). *Educating engineers for future industrial revolutions: Proceedings of the 23rd International Conference on Interactive Collaborative Learning* (ICL2020), Volume 2 (Vol. 1329). Springer Nature.

Ball, A. G., Zaugg, H., Davies, R. S., Tateishi, I., Parkinson, A. R., Jensen, C. G., & Magleby, S. R. (2012). Identification and validation of a set of global competencies for engineering students. *International Journal of Engineering Education, 28,* 156–168.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 1–22. https://doi.org/10.1080/10627197.2023.2202312

Cajander, Å., Daniels, M., McDermott, R., & B. Von Konsky (2011). *Assessing professional skills in engineering education, In J. Hamer & M. de Raadt (Eds.), Thirteenth Australasian Computing Education Conference* (ACE2011), Jan 17, 2011. Perth, Western Australia: Australian Computer Society.

Douglas, K. A., Neumann, K., & Oliveri, M. E. (2023). Contemporary approaches to assessment of engineering competencies for diverse learners. Johri, A. (Ed.). *International Handbook of Engineering Education Research.* (pp. 690–709). Routledge. https://doi.org/10.4324/9781003287483.

Evans, C. (2021, November 3). *Culturally sensitive, relevant, responsive, and sustaining assessment.* Center for Assessment. https://www.nciea.org/blog/culturally-responsive/culturally-sensitive-relevant-responsive-and-sustaining-assessment

Fajaryati, N., Budiyono, Akhyar, M., & Wiranto (2020). The employability skills needed to face the demands of work in the future: Systematic literature reviews. *Open Engineering, 10*(1), 595–603. https://doi.org/10.1515/eng-2020-0072

Fortus, D., Kubsch, M., Bielik, T., Krajcik, J., Lehavi, Y., Neumann, K., Nordine, J., Opitz, S., & Touitou, I. (2019). Systems, transfer, and fields: Evaluating a new approach to energy instruction. *Journal of Research in Science Teaching*, 56(10), 1341–1361. https://doi.org/10.1002/tea.21556

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116. https://doi.org/10.1177/0022487102053002003

Gay, G. (2013). Teaching to and through cultural diversity. *Curriculum Inquiry*, 53(1), 48–70. https://doi.org/10.1111/curi.12002

Geisinger, K. F. (2016). 21st century skills: What are they and how do we assess them? *Applied Measurement in Education, 29*(4), 245–249. https://doi.org/10.1080/08957347.2016.1209207

Gopen, G. D., & Swan, J. A. (1990). The Science of Scientific Writing. *American Scientist*, *78*(6), 550–558. http://www.jstor.org/stable/29774235

Gutiérrez Ortiz, F. J., Fitzpatrick, J. J., & Byrne, E. P. (2021). Development of contemporary engineering graduate attributes through open-ended problems and activities. *European Journal of Engineering Education*, 46(3), 441–456. https://doi.org/10.1080/03043797.2020.1803216

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. https://doi.org/10.1111/emip.12253

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, 67(3), 187–196. https://doi.org/10.2307/2668188

Hopkins, B. (2009). *Cultural differences and improving performance: How values and beliefs influence organizational performance*. Taylor & Francis Group.

Irish, R. (2015). *Writing in engineering: A brief guide*. Oxford University Press.

Jesiek, B.K., Zhu, Q., Woo, S. E., Thompson, J., & Mazzurco, A. (2014). Global engineering competency in context: Situations and behaviors, *Online Journal for Global Engineering Education*, *8*, 1, Article 1.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education, 104*(4), 454–496. https://doi.org/10.1002/jee.20104

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96*(3), 558–568. https://doi.org/10.1037/0022-0663.96.3.558

Kim, H., & Care, E. (2020). *Capturing 21st century skills: Analysis of assessment in selected sub-Saharan African countries* (p. 55). UNESCO Publishing.

Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice, 34*(3), 159–165. https://doi.org/10.1080/00405849509543675

Ladson-Billings, G. (1995b). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491. https://doi.org/10.3102/00028312032003465

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education, 67*(3), 268–279. https://doi.org/10.2307/2668195

Lohmann, J. R., Rollins, H. A., & Joseph Hoey, J. (2006). Defining, developing and assessing global competence in engineers. *European Journal of Engineering Education, 31*(1), 119–131. https://doi.org/10.1080/03043790500429906

Lyons, S., Oliveri, M. E., & Poe, M. (2025). *A framework for enacting equity aims in assessment use: A justice-oriented approach.* In Evans, C. & C. Taylor, (Eds.), *Culturally Responsive Assessment in Classrooms and Large-Scale Contexts: Theory, Research, and Practice* (pp. 88–105). Routledge.

Merzdorf, H. E., Jaison, D., Weaver, M. B., Linsey, J., Hammond, T., & Douglas, K. A. (2023). Sketching assessment in engineering education: A systematic literature review. *Journal of Engineering Education,* 1–22. https://doi.org/10.1002/jee.20560

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5–11.

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

Mislevy, R. J., Oliveri, M. E., Slomp, D., Cropped Eared Wolf, A., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment, (pp. 199–241). Bennett, R. E., Darling-Hammond, L. D., & Badrinarayan, A. (Eds.), *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy.* Routledge.

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment?. *Language Learning*, *59*, 249–267.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into Practice, 31*(2), 132–141. https://doi.org/10.1080/00405849209543534

Nasir, N. I. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of educational research*, *76*(4), 449–475.

National Academy of Engineering. (2004). *The engineer of 2020: Visions of engineering in the new century.* The National Academies Press. https://doi.org/10.17226/10999

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press. https://doi.org/10.17226/24783

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

O'Dwyer, E. P., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, *36*(3), 286–303. https://doi.org/10.1080/08957347.2023.2214652

Olds, B. M., Moskal, B. M. and Miller, R. L. (2005), Assessment in engineering education: Evolution, approaches and future collaborations. *Journal of Engineering Education, 94*, 13–25. https://doi.org/10.1002/j.2168-9830.2005.tb00826.x

Oliveri, M. E. (2019). Considerations for Designing Accessible Educational Scenario-Based Assessments for Multiple Populations: A Focus on Linguistic Complexity. cultural contexts and priorities in assessment [special issue]. *Frontiers in Education*, 4. https://doi.org/10.3389/feduc.2019.00088

Oliveri, M. E. (2020). Assessments used in higher education admissions, (pp. 233–236). In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective*. Cambridge University Press. https://doi.org/10.1017/9781108559607.025

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, *19*(3), 270–300. https://doi.org/10.1080/15305058.2018.1543308

Oliveri, M. E., & Markle, R. (2017). Continuing a culture of evidence: Expanding skills in higher education. *ETS Research Report Series*, *2017*(1), 1–8. http://dx.doi.org/10.1002/ets2.12137

Oliveri, M. E., Mislevy, R., & Elliot, N. (2020). After admissions: What comes next in higher education?, (pp. 347–375). In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective*. Cambridge University Press. https://doi.org/10.1017/9781108559607.019

Oliveri, M. E., Randall, J., Beck, M. F., & Poe, M. (2023). *Understanding Social Justice Features in Statistics Writing: A Corpus-Based Case Study of Two Undergraduate Statistics Courses*. (pp. 119–145). Brown, D. W., & Zawodny Wetzel, D. Corpora and Rhetorically Informed Text Analysis: The diverse applications of DocuScope. John Benjamins Publishing Company. https://doi.org/10.1075/scl.109.06oli

Oliveri, M. E., Slomp, D. H., Elliot, N., Rupp, A. A., Mislevy, R. J., Vezzu, M., Tackitt, A., Nastal, J., Phelps, J., & Osborn, M. (2021). Introduction: Meeting the challenges of workplace English communication in the 21st century. *The Journal of Writing Analytics, 5,* 1–33. https://doi.org/10.37514/JWA-J.2021.5.1.01

Oliveri, M. E., & Wendler, C. (2020). *Higher Education Admission Practices: An International Perspective*. Cambridge University Press.

Paretti, M. C. (2006). Audience awareness: leveraging problem-based learning to teach workplace communication practices, *IEEE Transactions on Professional Communication*, *49*(2),189–198. https://doi.org/10.1109/TPC.2006.875083

Paretti, M. C., & McNair, L. D. (2008). Introduction to the Special Issue on Communication in Engineering Curricula: Mapping the Landscape, *IEEE Transactions on Professional Communication*, *51*(3), 238–241. https://doi.org/10.1109/TPC.2008.2001255

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831–841. https://doi.org/10.1002/tea.21032

Pellegrino, J. W., DiBello, L. V., & Brophy, S. P. (2014). The science and design of assessment in engineering education. *Cambridge handbook of engineering education research*, 571–598. https://doi.org/10.1017/CBO9781139013451.036

Poe, M., Oliveri, M.E., & Elliot, N. (2023). The *Standards* Will Never Be Enough: A Racial Justice Extension, *Applied Measurement in Education, 36*(3), 193–215. https://doi.org/10.1080/08957347.2023.2214656

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

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice- oriented critical antiracist lens. *Educational Measurement: Issues & Practice, 40*(4), 82–90. https://doi.org/10.1111/emip.12429

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment, 27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Roschelle, J., & Penuel, N. (2006, June 27-July 1). Co-design of innovations with teachers: Definition and Dynamics (Conference session). In S. Barab, K. Hay, & D. Hickey (Eds.), *Making a Difference: Volume 2: The Proceedings of the Seventh International Conference on the Learning Sciences*. Bloomington, IN, 606–612.

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.

Steiss, J., Wang, J., Kim, Y.-S. G., & Booth Olson, C. (2024). U.S. Secondary Students' Source-Based Argument Writing in History. *Written Communication, 41*(4), 693–725. https://doi.org/10.1177/07410883241263549

Subheesh, N. P., & Sethy, S. S. (2020). Learning through Assessment and Feedback Practices: A Critical Review of Engineering Education Settings. *Eurasia Journal of Mathematics, Science and Technology Education, 16*(3), em1829. https://doi.org/10.29333/ejmste/114157

Suárez-Álvarez, J., Oliveri, M.E., Zenisky, A. et al. (2024). Five key actions for redesigning adult skills assessments from learners, employees, and educators. *ZfW*. https://doi.org/10.1007/s40955-024-00288-8

UNESCO. (2016). *Education 2030: Incheon declaration and framework for action for the implementation of sustainable development goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000245656

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3–14.

Wylie, E. C., & Lyon, C. (2016). *Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice.* CCSSO.

VOLUME II | SECTION 2

Innovations in Practice–Tools and Methods Serving Learning

Game-Based Learning: A Design-Based Theory of Teaching-Learning-Assessment Systems

James Paul Gee

Abstract

This chapter presents Game-Based Learning (GBL) as an integrated teaching-learning-assessment system rather than merely a tool for content delivery. It argues that good games inherently combine teaching, learning, and assessment through their design principles, which align with research in learning sciences. The paper outlines 20 principles that constitute effective GBL, including empowered learners through meaningful choices, problem-solving opportunities, contextual understanding, and embedded assessment. Rather than focusing on memorizing isolated facts, GBL prioritizes problem-solving skills within meaningful contexts. The chapter reconceptualizes teachers as designers who create or implement systems using these principles, whether through games or other learning activities. It further argues that classrooms represent complex systems with emergent properties that cannot be adequately studied through traditional randomized controlled trials. Instead, the paper suggests approaches from complexity science—such as agent-based modeling, network analysis, and systems thinking—can better capture classroom dynamics. While empirical research shows positive effects of GBL on learning outcomes, the author emphasizes that context remains paramount in determining effectiveness. The chapter concludes that an integrated approach to teaching, learning, and assessment is necessary to create learning experiences that engage students in the same way good games engage players.

Introduction

In this chapter, I develop an approach to Game-Based Learning (GBL) in which teaching, learning, and assessment are convergent, entangled, and inseparable. GBL on this view is a teaching-learning-assessment system that can be introduced into classrooms. In the end, I argue that classrooms are complex systems in the technical sense in which physicists use the term and, thus, not readily researched or assessed by controlled studies.

In earlier work (Gee, 2003), I claimed "Good video games are good for learning." What I should have said is, "Good games are good for teaching, learning, and assessment." This claim cannot be tested unless we know what "good" in "good games" and "good for teaching, learning, and assessment" means. Explicating the meaning of "good" here involved offering a theory about games and learning. And, then, when we are testing the claim, we are testing the theory, not some simple "fact."

Learning

Let's start with "learning." The theory of teaching, learning, and assessment in many schools today is based on what I have called the "content fetish" (Gee, 2004). The content fetish is the view that any academic area (whether physics, sociology, or history) is composed of a set of facts or a body of information and, thus, that the way learning should work is through teaching and testing such facts and information. Such learning can lead to students passing tests, but the information is poorly retained past the test and into life (Murre & Dros, 2015). Students can know lots of facts about physics, for example, but still be unable to solve problems in physics (Chi, Feltovich, & Glaser, 1981).

However, any actual domain of knowledge, academic or not, is first and foremost a set of activities (special ways of acting and interacting so as to produce and use knowledge) and experiences (special ways of seeing, valuing, and being in the world). Physicists *do* physics. They *talk* physics. And when they are being physicists, they *see* and *value* the world in a different way than do non-physicists (Glenberg & Gallese, 2012; Latour, 1999; Pickering, 1995). The same goes for gardeners, gamers, musicians, and mathematicians (Gee, 2011).

Problem solving is a much better goal for education than learning/memorizing facts. When people learn to solve problems, they use facts and information, along with other skills, to solve the problems. In the act, they both learn facts and can

solve problems, and they retain the facts much longer (Shaffer, 2007). We live in a world replete with serious problems. Learning to solve problems—which involves learning to make good choices (Swartz & Arena, 2013)—is crucial for individuals and society. Good games—and I will move next to saying what that term means here—are based on problem solving, not facts and information, though you need to learn and use facts and information to solve the problems in the game.

Good Games

For my purposes (Gee, 2003, 2004, 2008, 2013; Gee & Hayes, 2011; Gee & Shaffer, 2010), "good games" are games which are focused on problem solving and which use a specific set of design principles to teach people how to solve problems. These design principles are supported by research in the learning sciences, but they have often also been discovered by game designers who will go broke if people cannot learn (and enjoy learning) to play their often long and complex games.

Good games, first and foremost, honor the principle that a game's "game mechanics" must be well married/matched to the sorts of problems the game involves solving. A game's game mechanics are the actions and tools gamers use to solve the game's problems. The game mechanics must work well and powerfully—and be motivating to use—to facilitate solving the problems the game is about. If the marriage of game mechanics and problem solving, in this sense, is not good, then all bets are off about the game being a "good game."

Following this principle, there are other design principles that constitute a form of "baked in" good teaching, learning, and assessment. These principles can be stated in various ways and others would modify them in various ways. This is not a definitive list, but an example of what a design system for teaching, learning, and assessment as convergent might look like. Below are some of these principles. There are 18 principles grouped by "empowered leaners," "problem solving," "understanding," and "assessment." There are others not discussed here.

I. Empowered Learners

1. Choices

Good games make players feel like producers and not just consumers. Players' choices and actions in the game make the game world unfold and change, thereby becoming an important part both of the story in the game and the player's own story of what it meant to play the game.

2. Different Ways to Solve a Problem

Players need to be able to try different ways to solve problems and to find new ways to solve them when their problem solving gets too routine. Good games allow players to solve problems in different ways and to try new approaches. This allows players to see problems as part of a larger problem space.

3. Identity

Deep learning requires an extended commitment and such a commitment is powerfully recruited when people take on a new identity they value and in which they become heavily invested—whether this be a child "being a scientist doing science" in a classroom or an adult taking on a new role at work. Good games often offer players avatars—and sometimes let players design their own avatars—that exemplify and reflect an identity (with concomitant values) that the player's choices will modify and concretize.

4. Action

Cognitive research (Barsalou 1991a, b; Glenberg 1997; Glenberg & Gallese 2012) has argued that humans think and learn best when they have an action to take whose consequences they care about and when they are helped to be able to assess the results of these actions and to use these results as feedback about how to proceed. Good games motivate such actions and give players feedback about results and consequences and about alternative ways to proceed toward their goal.

5. Affinity Spaces

Good games do not just recruit software to teach and create problem solving. They very often also incorporate what I have called "affinity spaces." These are internet sites or places in the real world where gamers engage in social learning around strategies and problem solving and sometimes "mod" (modify) the

software of the games they play. And, of course, players often socialize within their games and engage in multi-player play where competition and collaboration intermingle. Thus, GBL often involves the combination of the game and affinity spaces, as well as socialization and collaboration within the game. This whole system is what I have called the "Big G Game," the Game as a teaching-learning-assessment system and not just the game as software. (https://home.edweb.net/big-g-game-based-learning/)

II. Problem Solving

6. Performance Before Competence

Good games use the principle of "performance before competence." They do not demand that players learn everything before engaging in action, since they want players to learn by doing and reflecting on what they are doing.

7. Time is Not the Measure of Learning

Unlike schools, "good games" do not usually use time as a measure of learning. It does not matter how long it takes a player to finish a level or a game or how many times they must play a level or the game to achieve mastery. Sometimes, those who take longer learn more. And mitigating the role of time means that the game does not discriminate against players who have come with less preparation than other players. In the real world, no one cares how long it took someone to learn physics when they win the Nobel Prize.

8. Information "On Demand" and "Just in Time"

Human beings are quite poor at using verbal information (i.e., words) when given lots of it out of context and before they can see how it applies in actual situations. They use verbal information best when a small amount is given "just in time" when they can soon put it to use and test their understanding of it. Larger blocks of information are given "on demand" when players feel they need it and are motivated to seek it out. This is how good games deal with information.

9. Well-Ordered Problems

Given human creativity, if learners face problems early on that are too free form or too complex, they often form creative hypotheses about how to solve these problems, but hypotheses that do not work well for later problems (even for

simpler ones, let alone harder ones). They have been sent down a "garden path." The problems learners face early on are crucial and should be well-designed to lead them to hypotheses that work well, not just on these problems, but as aspects of the solutions of later, harder problems, as well. Problems in good games are well ordered. In particular, early problems are designed to lead players to form good guesses about how to proceed when they face harder problems later on in the game. In this sense, earlier parts of a good game are always looking forward to later parts.

10. Cycles of Expertise

Expertise is formed in any area by repeated cycles of learners practicing skills until they are automatic (no longer require conscious reflection), then having those skills fail in ways that cause learners to have to think again and learn anew (Bereiter & Scardamalia 1993). In many games, at each level of the game, players practice a new skill set until they have achieved an automatic level of mastery. Then, they are confronted with a "boss battle" that demands a display of this mastery, but in ways that require some innovation and renewed conscious reflection on what they have previously learned and automatized. This is why players do not expect to beat the boss on the first or even the first few tries. This process gets them ready ("preparation for future learning") to learn new skills or take old ones to new levels, which they will do on the next level of the game.

11. Regime of Competence

People learn best when they are given problems within their "regime of competence" (diSessa, 2000), but at its outer edge. This way the problems feel doable but challenging, a highly motivating state for humans (and often the source of the state of flow). This is why games are rarely easy. Players often enjoy hard games but they demand that they are "fair." "Fair" here means that players realize when they fail that it is their fault (and a source of learning) and the game is not rigged against them and that with more effort and reflection they can and will succeed.

12. Fish Tanks (Models)

In the real world, a fish tank can be a simplified eco-system that clearly displays some critical variables and their interactions that are otherwise obscured in the highly complex eco-system in the real world. Using the term metaphorically, fish

tanks are good for learning: if we create simplified systems, stressing a few key variables and their interactions, learners who would otherwise be overwhelmed by a complex system get to see some basic relationships at work and take the first steps towards their eventual mastery of the real system (e.g., they begin to know what to pay attention to). Good games offer players fish tanks, either as tutorials or as early levels. Otherwise, it can be difficult for newcomers to understand the game as a whole system, since they often cannot see the forest because of the trees.

13. Lower the Consequences of Failure

Games often have levels or parts where the price of failure is lowered so that players are encouraged to take risks, try new things, and explore widely. Games alternate between spaces where players are encouraged to take their time to explore the lay of the land (the set of possibilities) and spaces where they are encouraged to use their growing knowledge of the system to make more rapid progress towards higher levels of skill. These two ways of learning have been called horizontal learning and vertical learning. Schools focus too much on vertical learning and too little on horizontal learning, yet it is horizontal learning that prepares learners to learn well in later vertical learning.

14. Skills as Strategies

There is a paradox involving skills: People do not like practicing skills out of context over and over again, since they find such skill practice meaningless, but, yet, without lots of skill practice they cannot really get any good at what they are trying to learn. People learn and practice skills best when they see a set of related skills as a strategy to accomplish goals they want to accomplish. In good games, players learn and practice (and, indeed, practice many times) skill packages as part and parcel of accomplishing things they need and want to accomplish. They see the skills first and foremost as a strategy for accomplishing a goal and only secondarily as a set of discrete skills.

III. Understanding

15. System Thinking

People learn skills, strategies, and ideas best when they see how they fit into an overall larger system to which they give meaning. In fact, any experience is enhanced when we understand how it fits into a larger meaningful whole. Players cannot just view games as "eye candy," but must learn to see each game (actually each genre of game) as a distinctive semiotic system affording and discouraging certain sorts of actions, interactions, and values. Good games help players see and understand how each of the elements in the game fit into the overall system of the game and its genre (type). Players get a feel for the "rules of the game"—that is, what works and what does not work, how things go or do not go, in this type of world.

16. Meaning as Action Image (Situated Meanings)

Learners need to learn to use both abstract and contextual meanings to think, reason, interpret, and solve problems. A word or concept like "democracy" has an abstract, categorial, definitional meaning, but takes on different shades and vectors of meaning in different specific contexts. For humans, words and concepts are most useful when they are clearly tied to perception and action in the world. Knowing what "abrasion" means in geology only in a definitional way is not as useful of knowing how it actually applies when you are doing geology in specific contexts. This is, of course, the heart and soul of video games. Even barely adequate games make the meanings of words and concepts clear through experiences the player has and activities the player carries out, not through lectures, talking heads, or generalities. Good games can achieve marvelous effects here, making even philosophical points concretely realized in image and action.

IV. Assessment

17. Assessment and Learning are Not Separate

We can readily claim that games are nothing but assessment, that in games, assessment and learning are the same thing. In games, players constantly choose, act, and get actional feedback from the game world (an assessment). Furthermore, after each action and across the game they must assess their own performance in order to get better and be able to finish the game. Games often give players tools to help them with these self-assessments. Finally, gamers often receive feedback socially through multi-player play and in interactions on affinity spaces. In good games, learning and assessment converge.

18. Stealth Assessment

Games often engage in "stealth assessment." The game collects (with players being aware of it) multifaceted information on the player's progress and can compare this progress both to the player's previous play and to a great number other players. This can allow the game to give players feedback about how they are progressing and how they compare to other players. The game can even adjust difficulty levels for different players or customize problems for them. The game can suggest what players should do next, given how they are progressing, and even encourage players to seek different or more innovative approaches to their problem solving.

19. Multifaceted Assessment

Good games do not just give players' grades—which offer little operational feedback—but multiple types of information, sometimes across time tracking progress, as well as sometimes comparison on each several different variables to how other types of players have progressed. This allows players to reflect on the data and form new strategies for getting better. And they can go to affinity spaces and learn about and share different strategies.

20. Assessment for Teachers/Designers

Much of the information about performance that game designers collect, in alpha and beta testing of their games and in collecting information about play styles over time, is used both to give feedback and encourage reflection on the part of players and of themselves as designers. Players also regularly give designers feedback via affinity spaces and other forums. Feedback is a two-way street. Designers use much of the information they collect on player performance—including the information players use to assess their own progress—to learn better how to do their job as designers (teachers, assessors).

Let me end this list by making it clear that the above principles are neither conservative or liberal, traditional or progressive. The progressives are right in that situated embodied experience is crucial. The traditionalists are right that learners cannot be left to their own devices, they need smart tools and, most importantly, they need good designers who guide and scaffold their learning (Kelly 2003). For games, these designers are good game designers. For schools, these designers are good teachers.

Game-Based Learning (GBL) as Design

For me, GBL does not mean using a game—though it can most certainly involve doing so—but using the design principles built into good games in or out of school. What we want to bring to in-school and out-of-school learning are teaching-learning-assessment systems that incorporate the sorts of design principles that good games use. These design principles can be used in many different activities and modes, not just games. This perspective casts teachers as designers.

Teachers in school should think and act like good game designers whether they are using games or other activities for learning. They should build good teaching, learning, and assessment principles into the games and other activities they use or ensure they have been already been built into the games and activities they use. They should create good social systems around the learning in their classrooms.

While using good games is something we should do, we should never use one tool (like a textbook) for everything and everyone. No one tool fits everyone. We want to network the best tools, activities, social systems, texts, technologies, and games together into a teaching-learning-assessment system. For example, here are just some of the tools ("game mechanics") students could use to learn about how pendulums work:

1. **Simulation Software** (Programs like PhET Interactive Simulations);
2. **Online Tutorials;**
3. **Lab Equipment** (Physical pendulums, stopwatches, protractors, and rulers for practical experiments to measure periods, lengths, and angles);
4. **Smartphone Apps** (There are apps that use the phone's sensors to measure periods and angles, turning the phone into a pendulum);
5. **Spreadsheet Software** (Tools like Microsoft Excel or Google Sheets to record data, create graphs, and analyze the relationship between variables like length, mass, and period);
6. **Data Logging Tools** (Devices that can record time intervals and angles with high precision for detailed analysis);

7. **Physics Forums and Online Communities; DIY Pendulum Kits: 3D Models and Animations** (Visual aids that help in understanding the motion of a pendulum in three dimensions);

8. **Augmented Reality** (AR) and Virtual Reality (VR): (Technologies that can simulate pendulum experiments in a virtual environment, providing an immersive learning experience);

9. **Mathematical Modeling Software** (Tools like MATLAB or Mathematica for more complex simulations and analyses of pendulum motion).

10. **Games** (Like Brain Pop's Pendulum Lab game).

These tools and others can constitute the "game mechanics" for different "levels" of understanding about pendulums. A system might well involve students choosing those tools which work best for them. Then our other design principles can be used to create a "teaching-learning-assessment system." The teacher is, then, designing in much the way good game designers do.

Empirical Research: Classrooms Are Complex Systems

There is a little commented on, but obvious, paradox in studying classrooms. Classrooms are complex systems in the sense in which physicists use the term. A complex system is a system composed of many components which may interact with each other. These systems are often characterized by the following features (Bar-Yam 2002; Ladyman & Wiesner 2020):

1. **Emergent behavior:** Complex systems exhibit properties that are not evident from the properties of the individual parts. The behavior of the system emerges from the interactions between its components and cannot be predicted by simply analyzing the components in isolation.

2. **Nonlinearity:** The interactions within a complex system are often nonlinear, meaning that small changes in input can lead to disproportionately large changes in output, and vice versa. This nonlinearity can lead to phenomena such as chaos and tipping points.

3. **Feedback loops:** Complex systems often have feedback mechanisms where the output of the system feeds back into the input, potentially amplifying or dampening effects within the system.

4. **Adaptation:** Many complex systems can adapt and evolve over time in response to changes in their environment. This is particularly true for biological and social systems.

5. **Self-organization:** Complex systems can exhibit self-organization, where order and structure arise from local interactions between parts of an initially disordered system, without external direction.

6. **Interconnectedness** and interdependence: The components of a complex system are interconnected and interdependent, meaning that the state or behavior of one component can significantly affect the state or behavior of others.

Examples of complex systems include ecosystems, the human brain, the climate system, social and economic systems, and many others. Classrooms, too, indeed, appear to fit these features well. A multitude of complex variables influence outcomes. These include student backgrounds, teacher practices, school culture, socio-economic factors, and many other variables. Since classrooms are composed of multiple people acting together—multiple physical and social brains using various tools and technologies all embedded within multiple complex institutions inside a highly diverse society—it is hard to see how they could fail to be a complex system.

Paradoxically, the gold standard of educational research—with government backing—has been randomized controlled trials (RCTs), where participants are randomly assigned to either a treatment or a control group (https://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp). This is just the method that will not work for complex systems.

Physicists study complex systems using a variety of methods and approaches. Some of these methods are mathematical modeling, computer simulations, network theory; nonlinear dynamics and chaos theory, big data and AI, and others. Because complex systems often involve phenomena that span different scientific

domains, physicists frequently collaborate with biologists, ecologists, social scientists, computer scientists, and others to study these systems.

The problem with the complexity of classrooms gets bigger when we realize that what we should be studying are teaching, learning, assessment, curricular, and social systems (all interacting), not isolated bits of them. Educators have sought alternative methods, such as mixed-methods research, quasi-experimental designs, implementation studies, design-based research, and even forms of A/B testing that engage in rapid cycles of testing and refining interventions in real-world settings, to learn what works and under what conditions. The problem is that in the mainstream of education and in the public view, controlled studies remain predominant despite large problems of ecological validity. Educators have not, for the most part, sought to study classrooms as outright complex systems in a truly interdisciplinary way.

There are, of course, educational researchers and cognitive scientists who have applied principles of complexity science to understand classroom dynamics (Gómez, Ruipérez-Valiente, & García Clemente 2022; Jacobson, Levin, & Kapur 2019; Keshavarz, Nutbeam, Rowling, & Khavarpour 2010; Knight 2022; Larson-Freeman 2016; Osberg & Biesta 2010). Agent-based modeling, network analysis, and systems thinking have been utilized to understand and explore classroom dynamics.

Agent-based modeling (ABM) is a powerful simulation tool that allows researchers to create computational models for studying complex systems. ABM is a computational approach used to simulate the actions and interactions of autonomous agents (individuals or collective entities such as organizations) with a view to assessing their effects on the system as a whole. It combines elements of game theory, complex systems, emergence, computational sociology, multi-agent systems, and evolutionary programming. In the context of classrooms, for example, ABM can be used to simulate the interactions between individual students (agents) and their environment, providing insights into how collective behaviors emerge from individual actions interactions (Jaiswal & Karabiyik 2022).

Network analysis examines the relationships and interactions among students and teachers to reveal patterns of communication, the flow of information, and the influence of social dynamics on learning outcomes.

This approach aligns with the principles of "systems thinking," which emphasizes the importance of understanding complex dynamic systems, including educational settings, as networks of interdependent components (Penuel, Sussex, Korbak, & Hoadley, 2006).

Systems thinking itself is a holistic approach that focuses on how different parts of a system interrelate and how systems work over time within the context of larger systems. In the classroom, systems thinking can help educators and researchers understand the complexities of educational processes, including how various elements such as student behavior, instructional methods, and curricular design interact to produce the overall educational experience (Abbott & Hadzikadic, 2017).

The application of complex systems theory to education can help in understanding how individual actions can lead to collective behaviors, how patterns of interaction affect learning, and how to design educational interventions that are sensitive to the complexities of real-world classrooms. Nonetheless, the studies in this area—and there are not a great many—remain outside the mainstream of research and policy in education and in education.

Studies

Since classrooms are complex systems, it is not surprising that controlled studies in education often reach rather mixed results. These studies rarely assess classrooms in terms of learning systems, but as discrete variables that can be controlled and isolated in the absence of any emergent properties. When we take them as a whole, we find out what every linguist already knows: Context is king. And context in the case of humans is itself a complex system.

There are, of course, important studies and meta-analyses that have contributed to our understanding of how games can be used for learning. For example, Clark, Tanner-Smith, and Killingsworth (2016) systematically reviewed research on digital games and learning for K–16 students and synthesized comparisons of game versus nongame conditions and comparisons of augmented games versus standard game designs. Arztmann, Hornstra, Jeuring and Kester (2023) examined the effects of games in STEM education. They found that game-based learning has positive effects on students' cognition, motivation, and behavior. This study also highlighted differences

based on certain students' background characteristics. Wang, Chen, Hwang, Guan, & Wang (2022) focused on the impact of digital game-based STEM education on students' learning outcomes across different STEM subjects. Their findings suggest that digital games are a promising pedagogical method in STEM education that effectively improves learning gains. Sitzmann's (2011) meta-analysis found that trainees using simulation games had higher self-efficacy, declarative knowledge, procedural knowledge, and retention of the material than trainees in more traditional learning methods.

Valerie Shute and her colleagues' seminal work on "stealth assessment" within games (Rahimi & Shute 2023; Shute & Becker 2010; Shute & Rahimi 2021; Shute, Wang, Greiff, Zhao, & Moore 2016) has explored how games can be used to assess learners' skills and knowledge in a way that is integrated with the gameplay, making the assessment process less intrusive and more engaging.

Farber (2018); Klopfer, Hass, Osterweil, and Rosenheck (2018); Isbister (2017); Plass, Mayer, and Homer (2020); Shaffer (2007); Squire (2011, 2021); and Toppo (2015) are among the best books on games for learning and cover among them a wide range of topics and approaches. These books all mix teaching, learning, and assessment with design principles. Each sees game design and curriculum design as similar activities when done right.

Conclusion

In this chapter I argued that teaching, learning, and assessment are co-dependent, interacting, convergent, and reciprocal aspects of a system that must be dealt with as a whole. GBL as a design theory offers one systematic theory of teaching-learning-assessment. Embedded in classrooms this system interacts with the classroom as a true complex system which has different emergent properties in different contexts.

In a good game players are always learning from teaching-learning-assessment principles "baked into" the design of the game by good game designers. That learning is based on choice, action, and problem solving and inherently involves assessment in several ways (e.g., feedback on performance, self-assessment, using evaluative information for reflection in and on action, assessment of a trajectory of progress, and comparison to other players and other strategies).

If a player plays a game like *Halo* on the hard difficulty level and completes the game, would you be tempted to give him or her a *Halo* test to assess their mastery? Surely not. The game is its own assessment. Why, then, don't our teaching-learning-assessment systems for algebra in school work the same way? Why do gamers like hard games and students don't like hard subjects in schools?

Reference

Abbott, R., & Hadžikadić, M. (2017). Complex adaptive systems, systems thinking, and agent-based modeling. In Hadžikadić, M., & Avdaković, S., Eds., *Advanced Technologies, Systems, and Applications. Lecture Notes in Networks and Systems*, vol 3. Springer.

Arztmann, M., Hornstra, L., Jeuring, J., & Kester, L. (2023). Effects of games in STEM education: a meta-analysis on the moderating role of student background characteristics. *Studies in Science Education*,59:1: 109–145.

Bar-Yam, Y. (2002). General Features of Complex Systems, *Encyclopedia of Life Support Systems*. EOLSS UNESCO Publishers.

Barsalou, L. W. (1999a). Language comprehension: archival memory or preparation for situated action. *Discourse Processes*, 28.1: 61–80.

Barsalou, L. W. (1999b). Perceptual symbol systems. *Behavioral and Brain Sciences* 22.4: 577–660.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise*. Open Court.

Buckley, J., Colosimo, L., Kantar, R., McCall, M., & Snow, E. (2021). Game-based assessment for education. In *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, OECD Publishing.

Davis, B., & Sumara, D. (2006). *Complexity and Education: Inquiries into Learning, Teaching, and Research*. Lawrence Erlbaum Associates.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science* 5.2: 121–152.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research* 86.1: 79–122.

DiSessa, A. A. (2000). *Changing minds: Computers, learning, and literacy*. Cambridge, MA: MIT Press.

Farber, M. (2018). *Game-based learning in action: How an expert affinity group teaches with games*. Peter Lang.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (2nd ed. 2007). Palgrave Macmillan.

Gee, J. P. (2004). *Situated language and learning: A critique of traditional schooling*. London: Routledge.

Gee, J. P. (2008). Game-like learning: An example of situated learning and implications for opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, L. J. Young, Eds., *Assessment, Equity, and Opportunity to Learn*. Cambridge University Press, pp. 200–221.

Gee, J. P. (2011). *Social linguistics and literacies: Ideology in discourses* (4th ed.). London: Taylor & Francis.

Gee, J. P. (2013). *Good video games + good learning: Collected essays on video games, learning, and literacy.* (2nd ed.) Lang.

Gee, J. P., & Hayes, E. R. (2011). *Language and learning in the digital age*. London: Routledge.

Gee, J. P., & Shafer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *EDge* (Phi Delta Kappa International) 6.1: 2–19.

Glenberg, A. M. (1997). What is memory for? *Behavioral and Brain Sciences*, 20.1: 1–55.

Glenberg, A. M., & Gallese, V. (2012). Action-based Language: A theory of language acquisition, comprehension, and production. *Cortex*, 48.7: 905–922.

Gómez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies*. 16.4: 500–515.

Isbister, K. (2017). *How Games Move Us: Emotion by Design*. MIT Press.

Jacobson, M. J., Levin, J. A., & Kapur, M. (2019). Education as a complex system: Conceptual and methodological implications. *Educational Researcher*, 48.2: 112–119.

Jaiswal, A., & Karabiyik, T. (2022). Characterizing undergraduate students' systems-thinking skills through agent-based modeling Simulation. *Sustainability* 14.19: 12817.

Kelley, A. E., Ed. (2003). Theme issue: The role of design in educational research, *Educational Researcher* 32.1: 3–37.

Keshavarz, N., Nutbeam, D., Rowling, L., & Khavarpour, F. (2010). Schools as social complex adaptive systems: a new way to understand the challenges of introducing the health promoting schools concept. *Soc Sci Med*. 70.10:1467–74.

Klopfer, E., Hass, J., Osterweil, S., & Rosenheck L (2018). *Resonant Games: Design Principles for Learning Games that Connect Hearts, Minds, and the Everyday*. MIT Press.

Knight, B. (2022). The classroom as a complex adaptive system (CAS): Credible framing, useful metaphor or mis-designation?. *International Journal of Complexity in Education*, 3.1. IntechOpen. https://doi.org/10.5772/intechopen.101699

Larson-Freeman, D. (2016). Classroom-oriented research from a complex system perspective. *Studies in Second Language Teaching and Learning*. 6.3: 377–393.

Ladyman, J., & Wiesner, K. (2020). *What Is a Complex System?* Yale University Press.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge. Harvard University Press.

Murre J. M. J., & Dros J (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS ONE*. https://doi.org/10.7: e0120644

Osberg, D., & Biesta, Eds. (2010). *Complexity Theory and the Politics of Education*. Brill.

Penuel, W. R., Sussex, W., Korbak, C., & Hoadley, C. (2006). Investigating the potential of using social network analysis in educational evaluation. *American Journal of Evaluation* 27.4:437–451.

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press.

Plass, J. L., Mayer, R. E., & Homer, B. D. (Eds.). (2020). *Handbook of game-based learning*. MIT Press.

Rahimi, S., & Shute, V. J. (2023). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development,* 125. https://doi.org/10.1007/s11423-023-10232-1

Shaffer, D. W. (2007). *How Computer Games Help Children Learn*. Palgrave Macmillan.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics educational game. *Computers in Human Behavior* 116: 1–13.

Shute, V. J., & Becker, B. J. (Eds.). (2010). *Innovative assessment for the 21st century: Supporting educational needs*. Springer-Verlag.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior* 63: 106–117.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology* 64.2: 489–528.

Squire (2011). *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*. Teachers College Press.

Squire, K. (2021). *Making Games for Impact*. MIT Press.

Swartz, D. L., & Arena, D. (2013). *Measuring What Matters Most: Choice-Based Assessments for the Digital Age*. MIT Press.

Toppo, G. (2015). *The Game Believes in You: How Digital Play Can Make Our Kids Smarter.* St. Martin's Press.

Wang, L. H., Chen, B., Hwang, G. J., Guan, J. Q., & Wang, Y. Q. (2022). Effects of digital game-based STEM education on students' learning achievement: a meta-analysis. *IJ STEM Ed* 9.26.

The Educative/Learning Portfolio: Towards Educative Assessment in the Service of Human Learning

Carol Bonilla Bowman and Edmund W. Gordon

Abstract

The Educative/Learning Portfolio: Towards Educative Assessment in the Service of Human Learning argues that our primary concern with portfolios should be their learning potential. The chapter documents generation, collection, and interpretation of pedagogically relevant evidence in search of student understanding. The portfolio processes of reflection, relationality, personalization and memorialization support enhanced motivation, engagement, effort, and metacognition. Portfolio learning is characterized by a caring, relational, and reflexive learning culture. The contents of the portfolio also provide a trove of in-vivo data that is then available for psychometricians and others, to distill for a variety of purposes.

Introduction and Rationale

In this chapter, we illustrate and elaborate on the concept of assessment in the service of learning. We propose a refocusing of our intention and vision of student portfolios from their use as an assessment tool (assessment portfolios) towards their intentional development as learning tools (educative portfolios). Assessment "in the service of learning" is a powerful concept, mandating one consider the assessment's learning value to the student first and foremost. Imagine how different our educational actions and instruments would be if it were required that all assessment must first benefit the learner, with the student learning "in vivo," and not simply demonstrating the knowledge test-makers have decided they should know.

The **educative** portfolio supports student learning throughout the portfolio-building process, with resulting portfolio artifacts (curated student work, student reflections, interviews, and other media) providing more useful and abundant evidence of achievement than a simple metric. The portfolio provides a window into the processes of the student's learning and offers appropriate contextual information for that learning. Let us contrast the analytic value gained through portfolios rather than placing standardized student test results under a microscope to be observed in static isolation from its context.

Portfolio artifacts lend themselves to better informing teaching and learning as they provide the contextual information, longitudinal examples, and in-depth reflective essays offering richer and deeper analysis of the student's learning. We reiterate the imperative that educational assessment should serve student learning, as we re-examine the educative value of portfolios. We reference Principle 3 of this series, " Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition."

The Educative Portfolio Approach to Teaching, Learning, and their Assessment

The portfolio is an intentional collection of student products/artifacts that represent their learning. Artifacts may be the results of teacher-initiated learning, group projects, and projects learners pursue independently. From inception, the primary focus of the educative portfolio is to benefit student learning. This liberating, mediating learning tool is guided by a framework of education values, habits of mind, and processes unique to their context. These portfolio contents and goals

are defined via teacher collaboration and intentionally designed to encourage active learning via exploration and experimentation. Teachers act as mentors and guides, providing challenges and direction, but the portfolio product truly belongs to the student. Student input is welcomed, encouraging engagement and investment in the portfolio process.

The educative portfolio is dynamic and open-ended without the need for a unitary model of excellence. Students are encouraged to experience the wonder of learning, following their curiosity and navigating their chosen paths of exploration and learning. How does one "measure" wonder, the experience of gazing upon a Picasso painting or the moving images created by a kaleidoscope? The educative portfolio culture focuses on students asking questions and discovering their answers rather than simply answering other people's questions. Their analysis provides rich material to support the diagnosis of learning and teaching deficits and remediation recommendations. Portfolio artifacts will also remain available for continuing meaning-making, examination, reflexion, and pedagogical analysis. While educative portfolio reviews may evaluate an individual learner's strengths, weaknesses, and progress, large-scale efforts to use comparative scoring across student populations has not proved valid, reliable, or useful.

Shouldn't young people experience a period of exploration, choice, and opportunities to freely express themselves? How can we teach the values of liberty when our education system measures a student's worth primarily by their ability to answer other people's questions—while teachers' jobs sometimes depend on their students' test scores? The educative portfolio experience is a learning process that extends beyond the limitations of space and form, reshaping what we can imagine education to be. In this process, the heuristic role of teachers enables students to "find out for themselves" within an environment of nurturance, support, and mutual respect, fostering a true teacher–student partnership.

Our Exemplars

While there have been many projects and studies of portfolios designed and used for the purpose of assessment, a paucity of portfolios created to enable and nurture student learning required us to pick extant models closest to our vision for the educative portfolio. Examples include two Middle School Studies (the Portfolio Assessment Collaboratives in Education [PACE] study and the Hernandez Bilingual School Portfolio Study). The third exemplar is the Ramapo College Teacher

Education Portfolio for undergraduate teacher education students. Though these projects were conceived for assessment, we found the richness of student portfolio work impressive and offered many opportunities to better understand student learning above and beyond questions we would even think to ask. When analyzed, these portfolios delivered an authentic, in-depth picture of what, why, and how students were learning and a richly nuanced picture of each student's learning process.

We integrate exemplar portfolio artifacts into this chapter to bring student voices into the discussion and extract the causative learning elements that supported their learning. Our first portfolio exemplar excerpts come from the Hernandez Bilingual School (HBS) during its first year of portfolio education. The Hernandez Portfolio Table of Contents/Checklist helps students organize the required artifacts in the portfolio and ensures that each is present. This bilingual portfolio project requires artifacts in both the student's first and second languages.

Hernandez Portfolio Content Checklist and Assessment Guide

Tables 1 and 2 help illustrate what a full portfolio looks like. The Hernandez Portfolio Content Checklist (Table 1) specifies what each portfolio should contain: an introductory letter explaining the purpose of the portfolio, the number of assignments across a range of disciplines and genres, explicit performance standards for included artifacts, explicit criteria for the whole portfolio, and an explanation of included artifacts learning units. The following specific artifacts are required: illustrations of group work; a range of perspectives on the student's growth and achievement including self-reflection, peer, teacher, and parent reflections. Multiple drafts permit observation of students' learning processes, growth, context, and representation of home culture and language. Students must check off that each of these components is present to submit their portfolio (See Table 1).

Table 1.
Hernandez Portfolio Content Checklist

Content Checklist	L1	L2
Introductory letter (should include student's understanding of why she does portfolios)		
Explanation of project or assignment for each entry (attached to work)		
Multiple drafts of work		
Collaborative piece		
Varied perspectives on student growth and achievement 1. Students reflections (attached to work 2. Peer reflections 3. Teacher reflections and comments 4. Parent reflections		
Explicit performance standards (rubrics, benchmarks)		
Range of disciplines represented by work		
Range of genres		
Selection of work that permits observation of growth		
Representation of home culture		
Number of pieces of work in each language		
**L1 Refers to the student's dominant language, L2 to their second language.*		

The **Hernandez Narrative Portfolio Analysis Guide** (Table 2) is organized according to the nine guiding principles and related big ideas and processes based on "habits of mind," that the teachers hope to engender in their students. Teachers met weekly over several months to form a consensus in defining the guiding principles and ensuring they honored their ethnically diverse student population. One or two artifacts are required in each of the 9 categories. For the assessment guide, excerpts from the student's portfolio provide evidence of habits of mind, big

ideas, and processes from each category and documented by the reviewer. Table 2 includes a completed narrative portfolio analysis (the sample does not show a quantitative assessment). Commentary is included where deemed important. The last 2 columns of the table are marked to indicate P-Presence of work, or Q-Quality of work (this experiment in quantifying portfolio analysis was found ineffective in adding useful information to the teacher guiding the learner.)

Table 2.

Hernandez Narrative Portfolio Analysis Guide

Guiding Principles (Habits of Mind)		Big Ideas	Processes	P/Q*	
1	**empathy/ perspective taking**	**similarity and difference, contexts, values and valuing**	**communication** **transference**		
Community service: I still feel bad for people who don't have places to stay or food to eat. The hardest thing was there are other families out there without Thanksgiving meals. I was just glad that Father Jack would be choosing which family would get the food. Time Traveller: The book begins with the Time Traveller telling some men his ideas of time and space. The book isn't written from his point of view; it's written from the point of view of one of the men in the room. *Commentary: Writing: In her writing, Yara realistically represented the dialogue and point of view of each of her characters in context in a sophisticated and complex manner*					
2	**reflection**	**any/all the big ideas**	**conceptualization, analysis, synthesis, evaluation**		
Science reflection: I learned how to make an effective water filter. I don't know why it's important for me to know this. Maybe we'll have a Dirty water Emergency in the future. Then this information will come in handy.					
3	**service/ responsibility**	**values and valuing, contexts, interaction and constraints**	**Decision-making** **self-regulation**		
Community service: What affected me most was that I was actually helping out. It wasn't a little thing, like picking up a gum wrapper off the curb. The few cans of corn I put in the box actually helped a family have a better Thanksgiving.					

Table 2. (continued)

Guiding Principles (Habits of Mind)		Big Ideas	Processes	P/Q*	
4	**imagination/ creativity**	**reflection, perspective, experimentation.**	**conceptualization**		
Writing Assignment: The Wise Boy: Long ago, in a Mayan city in Guatemala, now called Palenque, there lived a young boy. This boy was only 13 years old, but he was as wise as the wisest old man in the town. Every day, people came to him with questions and every day, he answered all of them. *Commentary This principle is one of Yara's outstanding characteristics in all her writing and other work.*					
5	**curiosity and inquiry**	**cause and effect**	**information processing** **hypothesis generation and testing, collecting evidence**		
Manta Ray report: My driving question was "What is a manta ray? I wanted to find out everything I could about mantas Water Filter: My hypothesis was that pouring water through layers of rocks, dirt and gravel would clean it because water gets filtered when it goes through the earth.					
6	**skepticism/ respect for knowledge/ wisdom**	**cause and effect** **interaction, constraints**	**critical thinking**		
The Wise Boy: "Boy, you can't always be right," said the old man. "You are wise beyond your years [but] you are not the smartest person who ever lived. Remember this, if you try to please everyone, you will probably end up pleasing no one."					
7	**cooperation/ collaboration**				
Math: The activity wasn't hard. What was hard was trying to get my group to work. We started out O.K., but never finished our work Social Studies: we pretended to have a meeting where people were trying to decide whether or not to make the Boston harbor Islands a national park.					

Table 2. (continued)

Guiding Principles (Habits of Mind)		Big Ideas	Processes	P/Q*	
8	**appreciation of diversity**	**Similarity, differences contexts, systems** **interaction, constraints**	**communication** **evaluation**		
Boston Harbor islands meeting: The power was exercised positively. But Mr. NEve kept trying to take the power and use it negatively. Ultimately, no one person had all the power. That's how it is when you vote.					
9	**environmental awareness**	**Systems, cause and effect, continuity and change**	**quantification** **conceptualization**		
Boston Harbor islands meeting: I would have voted to make them a national park so that there would be people to make sure they were clean, there would be more jobs, and the wildlife there would be better protected. See "Curiosity and inquiry"-water filter					

All excerpts in Table 2 are the student's actual words from a piece of work or reflection on work. The analysis notes her outstanding imagination and creativity in her written work. We can see her ease with scientific concepts such as "guiding questions" and hypotheses. Note the frustration in group work where the members do not live up to their responsibilities. Her empathy is represented by her response to a service activity in a soup kitchen. She notes that in a group decision made by vote, not all power is held in one person's hand. Each entry by itself may not convey much information, but together they create a powerful set of patterns and a picture of this student's concerns, processes, values and skills.

Portfolio Culture

Portfolio learning flourishes in a well-supported environment and helps create and sustain that culture of learning. In a healthy portfolio culture, students are seen as active, important, and valued agents, not as passive receivers of information. Dialog-centered instruction alternates with small-group work, and student collaboration in non-teacher-directed projects is encouraged. The portfolio culture is characterized by a common spirit of cooperation, openness, support, reciprocal concern, self-criticism, and revisions. Full-scale portfolio learning encompasses project work, performance tasks, writing processes, and reflections. The educative portfolio culture is intended to nurture student intellective achievement, encourage intellectual engagement, and expand opportunities for holistic learning. Teachers conference with individual students to personalize support and offer specific suggestions to best support that student's growth. The teacher is a facilitator and mediator who guides students in their learning process in an atmosphere of mutual respect.

Activities occurring around the portfolio process are also part of the "portfolio culture." These include portfolio interviews, exhibitions, and events where students present their work to their parents and teachers as a focal point of their parent/teacher/student conferences. HBS has an expedition each semester that takes students out of the classroom for more experiential learning and engages much of the community. Illustrating the value of educative portfolio culture, we observed high levels of engagement, we saw students so engrossed in writing activities that they eagerly skipped recess to work on revisions. We found students seeking resources in local libraries for projects and seeking out subject matter experts on their chosen topics of inquiry.

In the educative portfolio culture, teachers work alongside their students as caring coaches, helping chart a personal path for each student and celebrating the learner's growth as the portfolio grows with commemorative artifacts. The portfolio gives students ownership of the process. Evidence of a student's learning are not dependent on abstract rating metrics, they are tangible achievement products they can hold in their hands. Educators and students identify many outcomes enabled by a portfolio culture including efficacy and agency, pride and ownership, organizational skills, developing achievable goals, and self-reflection.

The following interview illustrates some of these aspects of portfolio culture. The interview occurred during the authors' participation in Portfolio Assessment Collaboratives in Education (PACE) portfolio development project. Evers is an African American 6th grade student with a history of non-engagement in school. His parents are supportive and highly educated, and his two older brothers are considered "superstars." Evers' parents believe he may have a slight learning disability, but they have not consented to having him tested. Evers is extremely proud of his work this year and though he has completed his projects and shown progress in his learning, he has not reached the standard for his grade. Though Evers typically mumbles, he enjoyed learning speeches from Macbeth and holds a sophisticated understanding of the main ideas of the play. Below are excerpts from his portfolio interview where he discusses his essay on Macbeth.

Interviewer: Has anything been hard for you in creating this portfolio presentation?

Evers: Yes, the work. All the work. This is very tough work. It's hard to write this kind of work and still enjoy something. Cause when you get work like ____ and ____, it's very difficult, it's very interesting, and it has a lot of meaning.

Interviewer: Do you think that the meaning is something you can apply in your everyday life, did you learn something about life?

Evers: I learned that um' like don't be fool and don't try robbing and murder cause it's gonna backfire.

Evers also published an imaginative story about a boy named Tom, who loves elephants. While Tom is riding on an elephant with his friend, he hears the elephants telling each other jokes. Tom is so shocked that he falls off, but his elephant saves him from being crushed. The following passages express many of Evers' fears and frustrations including difficulties with school, doubts about his intelligence, and the value he puts on interpersonal relationships:

That's how their friendship began. They played and had lots of fun together. They did not realize what month it was. Thomas rode with the elephants and could have anything he wanted. He learned to have the power of an elephant and the wisdom, too. Then Thomas noticed that it was almost time for school to start again. The night before the first day of school he ran away and went to live with the elephants forever. After a while the boy was getting smarter and smarter.

Speaking with him, we learn he holds low expectations for himself and possibly this contributes to his diminished efforts at school. Completing and publishing his storybook shows his capacity to produce thoughtful, creative work that holds meaning for him. His story also reveals some of his perceived difficulties and attempts at solutions. His mother was extremely positive about the work he was bringing home and is pleased that he's now doing his work!

You don't know…we used to have to find him in the house to get him to go to school. Look for him wherever he was hiding. It's the first time he's been included in the social interaction and the academic interaction of the school. They used to kind of sit him in a corner and he developed a very hostile exterior so that no one would bother him…Evers' mother.

Both parents credit the school's culture for turning around their son's history of non-engagement with schoolwork. Portfolio cultures honor the knowledge that learning is a process and not a static achievement, and the culture demonstrates the arc of learning by accepting and supporting students at any point along the trajectory.

Images and Metaphors for the Educative Portfolio

Images and metaphors communicate beyond the strictly literary. Malia, an African American 6th grade student, stated in a reflection, "My portfolio box is like a rose. It's like all my work is the thorns, and I'm the flower reaching to the top to be the best I can be."

Like a mirror, the educative portfolio reflects each person's unique qualities, including their learning processes. The learner can more easily engage in self-reflection and reflexivity with this travelogue of tangible evidence showing where their learning began, where it is now, and where they are going. As a dynamic guide, the portfolio celebrates the student's memorialized learning achievements and milestones. Students may even view their portfolio as a reference library of their identity. Viewing the multiple dynamic components of the learning process, we can see the educative portfolio as a kaleidoscope, with each rotation realigning the components into different relationships. Using the kaleidoscope metaphor, we imagine portfolio artifacts as elements that may be reordered, repatterned, and juxtaposed with each turn. Using the collection of each students unique artifacts, educators can analyze, diagnose, and support the learner's development in one connected recurring cycle.

How the Educative Portfolio Supports Learning Activities

Human Activity: Learning from Experience, Prior Knowledge, and Its Representations

When we problematize a situation, we learn while working towards and discovering solutions. That humans learn from activity is the foundation of much educational design. Portfolios provide a container for activity as a labor-a-tory for active learning. Human activity and problem-solving are learning processes. The ancient Bhagavad Gita teaches that the nature of the human mind is constant activity, that what defines us as living beings is activity, and to be human is to be active, and to be active is to be learning. Vygotsky (1962), among many others, believed in the educative power of social activity as a foundation of learning theory. Active learning through social interaction includes mediation provided by teachers, by more advanced peers, or can be provided in the portfolio structure itself. Active learning through social interaction includes mediation provided by teachers, by more advanced peers, or can be provided in the portfolio structure itself. The capacity to learn from active experience is based on the progression from experience to symbol, then to developed ability. The integration of prior knowledge and its representations helps build the learner's cognitive structures.

Piaget, a strong proponent of the idea that we learn by doing, posits that a child's interaction with phenomena is a primary motor of development, putting activity at the center of learning and knowledge development, before perception and language (Ginsburg & Opper, 1988). To understand the functional and relational processes of human activity in this context, with development as intent, we must document that activity. Memorializations are the process of documentation for an educative portfolio, providing opportunities for the learner to transform physical interactions with phenomena into cognitive structures. Once memorialized in portfolio artifacts the experience is transformed into symbolic representation. We might look at a 7th grader's portfolio as a memorialization of the learning and accomplishments in that phase of the student's life. Activity theory suggests that encounters with activity result in learning, while reflexive review via memorialization further strengthens it. Educative portfolios provide students with opportunities to transform physical interactions with phenomena into cognitive structures and memorialized in their portfolio artifacts, transforming their experience into symbolic representation.

Documentation and Memorialization

Representations of learning can be for the purpose of "learning from." The educative portfolio documentation is intended to enhance student learning through the process of active creation. Documentation involves the symbolic representation of ideas, thoughts, and actions. This is a learning process, and the resulting documents/artifacts provide the necessary raw data for assessment and analytics.

Portfolio artifacts provide a record of what has been learned, including the circumstances and learning conditions by which the learning was enabled. Through dialogic pedagogy, nurturing curiosity, and "the asking of good questions," a student's educative portfolio memorializes dialectical, interactional, and relational processes. The intention of the educative portfolio documentation/memorialization is to enhance learning for the student. Documentation involves symbolic representation of ideas, thoughts, and actions. The products of these memorializing and processual activities provide a wealth of material for pedagogical analysis of learning, teaching, and assessment. Patterns of consistency and the systematicity of these associations provide a better understanding of the learning mechanisms and their meanings for the learner and this understanding better informs and improves learning, with summary judgments focused on progression. Another benefit is helping students understand their own metacognitive functioning and more effectively self-manage their learning processes. This aligns closely with Principle #4 of this series with student autonomy promoted in the educative portfolio allowing the learner to pursue learning based on their interests and encourages student engagement and ownership of their learning.

Most learning situations require the learner to access multiple data points simultaneously and concurrently to enable sense-making. Facilitation of learning and using prior knowledge to continue that learning process require a constant flow of abstract conceptualizations as the learner moves from one sense-making challenge to the next. Think of learning as a series of sense-making challenges with mental tasks in a continuous flow of automatic relationship discernment across numerous and changing connectives, many of them represented by abstract symbols. Student reflexivity is engaged when reflection is intended to result in action. The educative portfolio thus serves as the vehicle and container, with memorialized student work and student reflections on their work generating reflexive processes that support growth in learning and analysis. Reflexivity is engaged when student reflection is intended to result in action. Engagement in the

activity is of primary learning value, with memorialization reflexivity as the second most powerful determinant. Activity is internalized and transformed into mental structures, begetting symbolic representation.

Learning and Symbolization

Symbolic representation and its documentation in educative portfolio development gird the utilities of the portfolio as a learning tool. In animal life, the brain or the neural system is programmed by the activity of motivation and sensory phenomenon, animals learn stimulation forcing their limbic and sensory systems to act. To accommodate its environment, an organism's neural cells are organized so that learning is the programming of these neural cells. For human beings learning to be symbol users, symbols begin as tools, as instruments for doing things. Evolving into images and visual representations, symbols facilitate memory and further organize brain cells to achieve specific purposes. Thus, the symbolic memorialization features of the educative portfolio facilitate the learning process.

Most portfolio content is written language, making the discussion of symbolization especially important. This symbolic documentation provides raw data for further "learning from." Vygotsky (1987) suggested that oral language is a first-order symbolization of thinking and that writing and reading are second-order symbols or abstractions. The translation of thoughts first to oral language and then into written language is a complex cognitive task and writing creates a new awareness about the nature and powers of language. Mario Vargas Llosa (1991) here suggests the power of written language to initiate new learning:

Important knowledge about reality always comes out of [writing] … through a … transformation of reality by imagination and the use of words … When you succeed in creating something different out of … experience, you also achieve the possibility of communicating something that was not evident before … But you cannot plan this transmission of knowledge (Vargas Llosa, n.d., p. 79).

Bruner (1966) also placed great importance on the internalization of language as a cognitive activity, positing that written speech may bear the same relation to spoken speech as algebra bears to arithmetic. He sees language as an impetus or an engine to thought; when we are engaged with words, we are "led forward by them," and language itself affects the way we use our minds (a reflexive relationship) (1966), p.104). The centrality of writing as the mode of documentation

in portfolios reinforces the process of writing as representing and leading thought. The following exemplar demonstrates this capacity; that translation of concrete phenomenon into symbolic representation is useful as it permits manipulation serving the process of learning.

In a portfolio interview with a student who recently arrived from Vietnam, we discussed an entry he had written in English. As he tried to verbally reflect and interpret the meaning and context of his work, his thoughts went beyond his capacity to express them in English, creating noticeable frustration on his part. To alleviate his obstacle, the interviewer asked him to relate his thoughts in Vietnamese. The student spoke easily, seemingly expansive and articulate. The interviewer nodded enthusiastically while the student continued in Vietnamese. She then asked him to reprise this same discussion in English, and he did so eloquently and with little hesitation. The student spoke with pride about his portfolio; he appreciated the ability to share it with others and provide evidence of his developing mastery of English. As the interview concluded, the student exclaimed, "I had no idea that you spoke Vietnamese!" The interviewer responded, "I don't," much to the student's surprise!

In the process of language learning, or any learning, we are often unable to manipulate more than one high-level cognitive activity at a time; the capacity to think in one symbol system and speak synchronically in another was not yet available to him. Expressing his thoughts in his primary language first allowed him to manipulate the translation from one system of symbolization to another as a separate function. Whether thinking is represented in language as written in a portfolio or as a reflection in a spoken symbol system, the capacity of the mind to simplify cognitive processes through the manipulation of symbols facilitates higher-level learning.

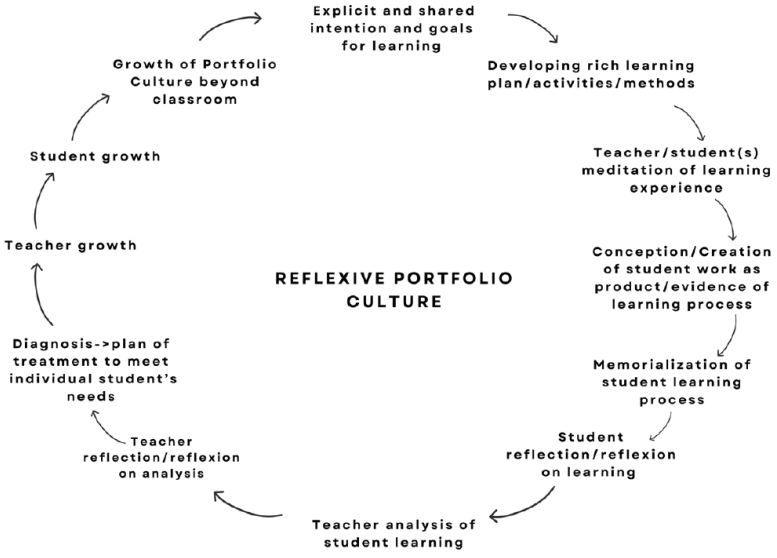
Analyses and judgments of the learning represented in portfolios are often based upon written content and drawn from writing's function as a reflection of cognitive processes. In the educative portfolio milieu, student writing assessment shifts from their degree of developed ability to a deeper analysis of what the writing teaches us about the students in ways that can enhance our ability to support and guide her growth. The focus is on what the writing indicates, requiring a more active analytical stance on the part of the teacher.

More About Reflectivity and Reflexivity: Learning from Documented Experience

Portfolios support and nurture a culture of reflective/reflexive teaching and learning, with memorialization artifacts representing learning for the purpose of "learning from." Earlier, we discussed the portfolio as a mirror, a reflection that allows observation of one's own processes and products. Reflection may be celebratory or commemorative, both worthy concepts. Reflexivity moves beyond reflectivity and implies intention in the review of one's own biases and reactions to their assessment and analysis of work. These insights can lead to more skillful actions and greater discernment in thinking. For example, if we reflect on an interaction with another person, our purpose may be to re-experience pleasure or to ensure we remember the details. If we review this interaction to critique or to deepen our understanding of our own perspective, this moves into the sphere of reflexion. When our review is intended to inform our future actions, such as whether to offer an apology or request more communication, this is now a fully reflexive process.

Reflexivity is a cyclical sequence (See Figure 1) in a continual process of internal investigation with the intent of improvement. Applying our kaleidoscope metaphor, we see that each turn of the lens provides a new picture of the same components. Every new piece of learning, like the flapping of a butterfly wing resounding across the universe, is rearranging and expanding our understanding and informing our actions. We see in the diagram below how reflexivity guides portfolio culture and brings depth to understanding our own and others' assumptions, perspectives a

Figure 1.
Reflexive Portfolio Culture Dynamics



What better use of the portfolio's potential than the pursuit of critical understanding, learning synthesis, analysis, and disconfirmation of assumptions—all aspect of reflexive thinking? Interpretation, imagination and reflexivity may be seen as mentation in its highest expression, and all three can be served by the documentation and memorization functions of the portfolio.

Reflection is intimately connected with how language structures our understanding of our experience and places it within a cognitive framework. Active involvement in reflective assessment encourages students' cognitive and metacognitive awareness of their learning process, providing a sense of control over the learning process, and encouraging self-determination in learning. Cognitive and metacognitive awareness of our learning process internalizes our locus of control and self-determination, shown to have positive effects on student achievement

(Pressley & Ghatala, 1990). In conventional disciplines of knowledge, we know concepts and theories serve to illuminate and connect observed facts. Portfolio reflections helps unify the learner's scatter of knowledge and refine how that knowledge relates to them as an individual.

Reflection is a critical part of teacher and student behavior in portfolio culture and fundamental to helping students develop work of higher quality, by reflecting on one's own work in relation to an external or internal set of criteria. Good teachers model such reflection in the substance of their own teaching. Through reflexive activity, the mechanics of analysis and appraisal become tools of teaching and learning. Reflective thinking becomes a practiced central part of student activity when students are expected to complete and save multiple drafts of works and incorporate their reflections into the final product. Additional reflection opportunities may include portfolio presentation days, exhibits in the classroom or for the community, or archives containing student work demonstrating fulfillment of portfolio criteria.

An exemplar: The Ramapo College Teacher Education Portfolio project focuses on self-reflection to support student development, and the portfolio is a requirement for teacher education graduation and teacher certification at the College. Self-reflection is intended to instill reflective and reflexive habits of mind in preparation for successful student teaching and candidates are encouraged to continue this practice. Most impressive about this portfolio process is the insight students gain about themselves as they collect, select, systematically review, and reflect upon their best work throughout the entire four-year program. This insight is illustrated in the following student reflection:

My portfolio helped me to see that my classes have helped me to be prepared in some ways. My portfolio showed me the different areas of teaching and why they are important to us as teachers. The assignments I have added to my portfolio, I feel really fulfill the requirement. I also like the fact the portfolio helps to show us our weaknesses and strengths when it comes to teaching. The reflections I have written throughout this portfolio have helped me realize what I need to work on as a teacher. I do have a few weaknesses that I feel need to be addressed before I student teach next spring, and I will work very hard to work in those areas and improve them. —Dave

In their portfolio interviews, Ramapo College students have remarked, "I didn't realize I did so much work," or "I am surprised that the work I did really makes me feel prepared to begin my student teaching," as they build their portfolios. Most satisfying as an educator was hearing, "I see now how all the classes fit together." Fine-tuning their self-perception as they move from college student to future teacher, their portfolios provide more accurate evidence-based perceptions. Those who began the program with trepidation gain confidence through their documented accomplishments and abilities and conversely may temper overly positive misconceptions about one's level of teacher competence.

Relational and Contextual Holistic Learning

An important argument in favor of educative portfolios is the integration of the cognitive, affective, and conative ways of knowing. Learners exist within a variety of global, community, classroom, and family ecosystems, that differ for each learner. As Moll (1990) argues, cultural context may imbue the learner with strengths that they may build upon. Educative portfolios are holistic and integrative, uniting various aspects of a learner's education experiences. The educative portfolio may illuminate the cultural context of the learner's life, as well as important cultural-historical contexts.

An exemplar of the portfolio's relational and contextual value may be seen in this predominantly Navajo School District in Chinle, Arizona, where sixteen 5th grade teachers participated in developing a portfolio assessment. Aligned with the Navajo-based curriculum and Arizona's curriculum frameworks, the project was intended to address the cultural context of Navajo students using criteria and rubrics reflective of their cultural values. For example, their Life Skills standards are interpreted in terms of Navajo philosophy, with well-being portrayed as the integration of mind, body, and spirit. Identified capacities were complex, such as cultural knowledge and self-direction, rather than simplistic facts and skills. The deforestation project the teachers designed asked students to work together and investigate a particular aspect of deforestation in their local area, one in which they would become an "expert."

Students gathered information from multiple sources, analyzed the problem, and proposed a solution for a final presentation (Farr & Trumbull, 1997). This holistic, integrated learning paired important, real-life cultural and economic issues with high-level intellective process as students were challenged to view the harm of

deforestation as it related to their own community. In their final presentation, students were directed to present their analysis of what steps could ameliorate harm to the environment and the people affected. Task structure encouraged students to use their knowledge to problem-solve in an authentic way. This included collaboration with peers, integrating various concerns and perspectives in their final recommendations, and sharing their findings with the community. Exemplifying holistic and contextual education, the deforestation project required students to have a relational and global understanding of the balance of nature as well as considerations of sustainability.

Requisites to Designing and Operationalizing Effective Educative Portfolios

Portfolio Curriculum and Pedagogy

> *"All curriculum plans are tentative, and children modify them by their response. Like the universe, curriculum is always expanding,"*
> —Jones, E. (1987).

Like the universe, educative portfolios expand the concept of curriculum far beyond "traipsing over trivia," and other content mastery tasks characterizing many curricula. Emergent Curriculum, as discussed in Jones' (1987) work, puts curriculum into the hands of the learner by integrating their interests, their questions, and their community knowledge. Traditionally, curriculum references a curated collection of knowledge and skills. This academic canon, divided by disciplines, and building through the grades, is intended to turn out "educated" children. That standardization of curricula is in part due to required symmetry between what students are taught and the secret content of high stakes, external assessments. In contrast, the educative portfolio is an instrument that supports education as the construction of one's own knowledge and meaning. Educative portfolios enable a curriculum attentive to developing an individual's mental capacities and heartful intelligence. Learning to develop effective strategies to think and solve problems should hold the primary position, with curriculum content serving as a beneficial tool.

An impressive example of an inquiry-based middle school curriculum with complex interdisciplinary projects comes from Saturno and Wolf (1997). This integrated archeology and mathematics project is also a prime example of "problematizing" learning. Saturno and Wolf identified "four core skills… at the heart of archeology: "1) genuine curiosity, 2) careful inquiry, 3) thoughtful inference, and 4) humility" (p.5). Students were directed to take the viewpoint of archeologists discovering hieroglyphs of the Mayan numerical system and try to decode their meaning. This activity was an "entryway" to study of the Mayan culture, arts, myths, and the scientific and mathematical knowledge of these ancient peoples.

This level of complex curricula is a requisite to supporting an effective and strong portfolio, providing ample opportunity for student self-reflection and useful analysis of learning. One student was so inspired by the task that they extended their inquiry into additional learnings about the Roman and other ancient numerical systems. The portfolio curriculum's flexible and open-ended format supported this student's research even further, as they began to study the Arabic numerical system and how it developed the sophisticated and original invaluable contribution of the zero placeholder. Our goal is to have educative portfolios embedded in this type of rich curriculum, supported by engaging pedagogies prioritizing holistic student development. We propose that developing students' abilities to apply their knowledge in solving novel problems will better prepare them for success in a constantly changing world.

Complex, authentic tasks are necessary to develop capacity and depth of reflection. Wiggins and McTighe (2005) defined a set of design prompts using six elements in creating such portfolio-worthy pedagogical "gems." Complex units of learning support an effective and strong portfolio while providing ample opportunities for self-reflection and useful analysis of learning. Educative portfolios enable a curriculum more attentive to student engagement and the development of genuine mental and heartful abilities. We reiterate that assessment should be a tool in the development of students' mental abilities rather than the primary driver of education. Educative portfolios and a curriculum prioritizing students' mental and heartful development offer a solution.

Curriculum Principles of Caring

Another proposed requisite is placing the evaluation of student learning into the hands of their teachers. Teachers are the experts on the contexts and needs of their students, families, and communities. When teachers are empowered to determine the specific values and goals to guide their curriculum and teaching practice they can provide symmetry between what is taught and what is assessed. Hernandez Bilingual School teachers met regularly over a few months to hammer out a consensus on the principles that would guide their portfolios. Together, they created an assessment guide respectful of the students' first language, knowledge, and skills alongside their acquired English language knowledge and skills. By creating portfolio frameworks locally, teachers were able to chart their path and build a cohesive portfolio program and culture.

Passionate, Caring, Reflective Teaching Persons and Pedagogy

Educative portfolio programs require greater teacher effort and time than needed for standardized test preparation and externally scored multiple-choice tests. In truth, this is often a stumbling block to establishing portfolio cultures more broadly. We must remain cognizant that without a conscientious commitment from schools and districts to provide the time and resources necessary for teachers to succeed, educative portfolios may burden already overburdened teachers. However, dedicated teachers such as those at Hernandez continue to strive and develop these tools to attain their desired teaching and student outcomes. Hernandez teachers feel it is worth their time and effort to create an experience providing their students with lasting positive life-changing impacts. Portfolio curriculum and pedagogy are designed to place responsibility on students to make decisions about their learning. The heuristic nature of the portfolio values questioning and allows students to follow their own compass through organic pathways of authentic learning. Guiding principles and frameworks place no boundaries around learning content or mode of representation. Instead, they support "the having of wonderful new ideas" (Duckworth, 2006), without squelching passion and independence in the learner. In a portfolio culture, pedagogy is mediative, and the curriculum is adaptive to the learner.

Customization and Personalization of Learning

The educative portfolio's adaptive structure meets the needs of both individual learners and the general developmental needs of the group. With teachers in a more heuristic role, tasks, and activities can be self-initiated, with students owning their learning. Each portfolio becomes a stable trove of information about that student's learning, illustrating their strengths, indicating where they need support, and providing the depth of information needed to inform analysis, diagnosis, and remediation.

In discussing teachers of excellence, Ladson-Billings sees teachers in the role of coaches, sharing learning responsibility with students. This shared responsibility for learning between teachers and student is an ideal model for the educative portfolio. Each student has a coach to ensure they meet their own expectations as well as external learning goals. Ladson-Billings also discusses "non-unitary" criteria for excellence, acknowledging that excellence has multiple valid forms. The educative portfolio orients towards student learning and encourages relational, dialogic pedagogy. Portfolio dialog is organically generated, with teachers, peers, parents, family, and community members external to the school brought into the conversation.

The educative portfolio challenges our enculturated conflation of learning with assessment. Minimizing the focus on standardization for comparison and prioritizing service to the student's learning experiences invites the paradigm shift discussed in this chapter. An extant model of personalization of learning is seen earlier in our nation's history, in one-room schoolhouses where personalization was a necessity. Julia Weber Gordon (1946) relates her teaching experience in such a school. Her book, The Country School Diary, illustrates a learning community where each student's curriculum addresses their personal and unique needs, passions, and context. Vibrant and detailed models like hers demand that we look backward as well as forward for inspiration.

Structuring a Memorialization of Students' Work

Some of the most brilliant scholarly work has been achieved when amorphous, rich, lived knowledge is systematized and structured, gathered into meaningful groupings that clarify and facilitate the transmission of meaning. Though the structure is adaptive, the educative portfolio still requires a framework to structure the contents and uncover patterns of meaning. A Ramapo teacher education

student reflects, "I also like the fact the portfolio helps to show us our weaknesses and strengths when it comes to teaching. The reflections I have written throughout this portfolio have helped me realize what I need to work on as a teacher." A well-structured portfolio framework brings patterns concretely to the fore allowing the student to analyze his/her own achievement.

The student reflection sheet below for "Habits of Mind: Problem Solving" is an example of a mediative heuristic guided portfolio reflection in middle school. Students were directed to select work illustrating their problem-solving ability. Lucia chose to reflect on a paper she had done on the 6th grade overnight field trip that included a problem with bugs.

1. Briefly explain the problem you were trying to solve.
   The bugs at the overnight field trip.

2. Explain how you went about solving the problem. What did you think about first? What steps did you take?
   I just wanted to go home. I just covered my face and I moved and tried to go where there wasn't no bugs.

3. Explain what you think is special about this assignment. Why does it stand out as a good example of problem-solving?
   Because then the bugs stopped coming to me, although they came back.

4. What was your biggest challenge in solving this problem and how did you deal with it?
   Them bugs up in my face, I couldn't deal with it. But then to solve the problem I stopped complaining,

5. What piece of advice would you give so someone who was working on a similar problem?
   Bring something that would keep the bugs away from you.—Lucia

Lucia shows wisdom beyond her years expressing that we can help solve a problem by letting go of complaining! Though Lucia's effort is incipient, the structured reflective and reflexive questioning helps her see, and perhaps better understand her thought processes framed in a unique, coherent structure.

Analysis Appraisal, Evaluation, Judgment, and Interpretation: Byproducts of the Portfolio

Educative portfolio creation encompasses an inventory of cognitive activities that facilitate learning including attention, engagement, discrimination, selection, choice, and comparison. The documented artifacts offer bountiful data for learners, teachers, and others to analyze, appraise, judge, and interpret for understanding, valuing, and supporting the learning of the student. We enhance student learning by creating tasks and student learning is further enhanced by their memorialization. We believe their analysis is best integrated into a reflexive cycle of learning growth when the teachers, who know their students best are the ones designing the tasks. The educative portfolio is a powerful tool for learning with analysis of that learning providing higher fidelity, in-depth, and more useful data to better support and direct that learning.

Scaffolding the Process of Analysis

How do we approach the analysis of student learning from portfolio artifacts? The qualitative assessment of portfolios at Hernandez demonstrates a reflective and reflexive approach. These four heuristic tools (Tables 1, 2, 3, and 4) are part of one whole integrated analysis process, the goal of which was to be systematic, equitable, and experiment with quantification. Their heuristic Narrative Portfolio Analysis Guide (Table 2) uses as criteria the 9 guiding principles and asks the reviewer to assess both the presence of the criteria and the quality of the criteria. Table 3, Qualitative Portfolio Review Guiding Questions, asks probing questions concerning the individual student's achievement and portfolio structure itself. It is a reflective/reflexive tool to probe the validity of claims in terms of sufficiency of evidence. The questions address the student's work, the student's understanding of their own learning, and the capacity of the portfolio to answer those questions.

Table 3.

Qualitative Portfolio Review Guiding Questions
A. Evidence of Growth Do the contents of the portfolio sufficiently allow for a meaningful analysis of the student's growth and achievement, multiple perspectives and non-unitary standards? If yes, explain and illustrate.
Commentary and evidence:
B. Evidence of Achievement
Depth Criteria: Does student work demonstrate understanding through application of concepts? Does student work demonstrate conceptual richness? Does student work demonstrate complex thinking? If yes, provide sample evidence.
Commentary and evidence
Breadth Criteria: Does student work demonstrate a breadth of conceptual understanding and skill? Does she/he make connections between various disciplines to solve problems? Is disciplinary knowledge presented in an integrated fashion? Describe. Does the student demonstrate ability to take multiple perspectives?
Commentary and evidence:

Table 3. (continued)

C. Multiple Perspectives: Are multiple readers and their perspectives included in the portfolios (such as peers, teachers, parents, and family members, experts in the field). Do these various perspectives reflect some common judgments? Provide evidence.
D. Non-Unitary Standards How does the student's work compare to the definitions of excellence held by peers, teachers, family/community? How does the student's work compare to the definitions of excellence being developed by the student?
E. Student Self-Review:
Commentary:
Teacher Analysis and Plan of Instruction:

In Table 4 we address a qualitative aspect of the assessment focusing on academic growth as well as level of achievement when compared to grade level expectations and standards. The Hernandez Achievement Summary for an individual student's achievement, translates the narrative analysis (Table 2) to a set of numbers. The numbers you see in Table 4 are totals taken from the narrative analysis guide (Table 2). Table 4's cumulative growth assessment score and level of achievement score are calculated from the evidence gathered in Table 2 (scores were not included).

Table 4.
Hernandez Achievement Summary

Hernandez Achievement Summary	1	2	3	4	5
A. Growth in academic proficiency from the beginning of the year	|	2	|	|	|
Evidence: Yara started out very advanced, but it appears that major progress has not been made. As this is the first year of the portfolio program we do not have anything to compare across years.					
Strengths and things to work on: Yara uses the draft and revision process effectively. Changes that are more than mechanical often occur. Her level of skill in writing puts her at the top of her class. I feel she needs more advanced challenges to grow.					
B. Achievement (relative to absolute standards) overall	|	|	|	4	|
1. depth	|	|	|	4	|
2. breadth	|	|	3	|	|
Yara's growth achievement is 2 and her achievement overall is 4. This fraction, 2/4, represents the effectiveness of the school's program in supporting learning. Any number under 1 means that the student has not been challenged to their potential.					

How can such a tool be useful? In a summary analysis of Yara's Growth in Academic Proficiency, one reviewer notes that since Yara's artifacts show excellent advanced work from the start of the year it was difficult to see progress though her level of achievement was high across the whole year. This suggested that a higher-level scaffolding may be needed to provide the challenge and support necessary for her to fulfill her potential. This narrative evidence provides insight into individual needs of students and allows teachers to better meet their students' needs. Yara's need for higher level instruction and challenge could easily slip through the cracks if we relied on one static and unitary metric of her level of achievement relative to other students her age and grade.

In the final analysis, the first author believes that the quantitative element did not add information above that which was already apparent in the qualitative analysis of portfolio artifacts. And, all that adding and dividing took a lot of time that could be better spent focusing on the teaching and learning use of the portfolio. Moving forward with analysis and learning analytics seems the more promising route to pulling out evidence that will put learning first.

The Educative Portfolio Facilitates Program Review

The chapter's first author also participated in developing the Teacher Education Portfolio assessment for Ramapo College's pre-service teachers. They began with the contention that to develop competent learners and teachers they must evaluate not just students' acquired knowledge but also their processes and future ability to apply their knowledge in the teaching classroom. They developed the portfolio to assess pre-service teachers as well as accountability of the Teacher Education program. Their goals included that: (a) students become more reflective and reflexive practitioners; (b) students better understand the standards of teaching practice; and (c) students better understand their strengths and weaknesses. Portfolio completion is intended to help students integrate and internalize their pedagogical learning experiences and develop reflective and reflexive habits that enable modification of practices towards more effective teaching going forward.

Benchmark trainings helped faculty establish common acceptable interpretations of defined criteria using the New Jersey Professional Teaching Standards (NJPTS) as a base. Aggregating data across a selected range of students allowed evaluation of the program's strengths and weaknesses and suggested necessary program changes. Given the relatively low stakes for individual students and the small local program, we were able to mitigate difficulties found by larger entities (Koretz et al., 1993) in creating a valid and reliable portfolio learning and assessment system.

Conclusion: The Educative Portfolio as Instrumental to Human Learning.

More than 10 years ago, the Gordon Commission on the Future of Educational Assessment concluded that current academic assessment practices were deficient for students in this 21st century (Gordon Commission, 2013). The education scholars of the Gordon Seminar on Educational Assessment in the Service of Learning are currently exploring possible solutions. Though convincing arguments have been made that a technology designed to determine status cannot readily

deliver effective learning interventions (Gordon Commission, 2013), we believe that by re-intentioning assessment portfolios towards educative portfolios, they can readily deliver the effective intervention solutions needed to enhance student learning. We see the educative learning portfolio as the best vehicle to assist learners in transforming non-static events (like activity and experience) into abstract theoretical constructs and symbols.

In this chapter, we reference the educative portfolio as a way of actualizing and refocusing our energies toward assessment that is in the service of learning. We need to develop greater capacities to engage in these broad analyses to better serve our learners. The kaleidoscopic metaphor reflects both shifting a paradigm using familiar components and our understanding of the dynamism of the learning process. We assert that the kaleidoscopic-like symbolic images of thinking processes can be better recognized once memorialized in portfolio artifacts. By creating new relationships with each turn of this kaleidoscope we can advance from assessing a learner's static state towards the dynamic, conceptual understanding available given the breadth and depth of data evidence provided through the educative portfolio.

A second concern is humanistic, in knowing the harm being done to our children by making standardized measurement the most powerful driver of our educational system. How can we best love our children/students? The institutions we created for schooling and assessment can provide a more fertile ground for learning if we prioritize a more relational, caring, personalized, and holistic education. Our current educational approach does not reflect what is known about the power of love and caring for successful learning, as well as the value and power of healthy human relationships. The "team" effort in completing a portfolio has teachers and students working shoulder to shoulder—as partners. The final portfolio interview or presentation is not an exam, it is an opportunity for the student to shine. Family and others may be invited for the presentation and a celebration may even be included. As human animals, we retain a strong drive to belong, to be loved, and to learn. When we put all of those together, we harness the power of these drives towards positive outcomes—so much wiser than working against them.

We continue to amass knowledge about what best enables learning, but how do we implement it? With technological advances allowing us to examine and analyze large amounts of qualitative data more effectively, portfolio pedagogy can now

move forward with analytical support applied to this more holistic learning. Student portfolios are ready to mature into their more natural and comfortable role as powerful learning tools and support enhanced student learning and pedagogical analysis. With the many learning opportunities educative portfolios offer and the evidence they provide for interpretation and analysis, the educative portfolio clearly shows its value as a powerful instrument in the cultivation of human learning.

References

Adams, D. H., Wilson, S., Heavy Head, R., & Gordon, E. W. (2015). *Ceremony at a Boundary Fire: A story of Indigenist Knowledge.*

Bowman, C. S. B. (1999). *Portfolio assessment in a bilingual/bicultural middle school: Building upon the PACE model* (Doctoral dissertation, Teachers College, Columbia University).

Bruner, J. (1966). *Toward a theory of instruction.* Harvard University Press.

Duckworth, E. (2006). *The having of wonderful ideas and other essays.* Teachers College Press.

Farr, B. P., & Trumbull, E. (1997). *Assessment alternatives for diverse classrooms.* Christopher-Gordon Publishers.

Freire, P. (2018). *Pedagogy of the oppressed* (50th anniversary ed.). Bloomsbury Academic. (Original work published 1968).

Gardner, H., & Hatch, T. (1989). Educational implications of the theory of multiple intelligences. *Educational Researcher*, *18*(8), 4–10.

Geertz, C. (1974). From the native's point of view: On the nature of anthropological understanding. *Bulletin of the American Academy of Arts and Sciences*, *28*(1), 26–45. In K. Gergen (2009). Relational being: Beyond self and community. Oxford University Press.

Ginsburg, H. P., & Opper, S. (1988). *Piaget's theory of intellectual development* (3rd ed.). Prentice-Hall, Inc.

Gordon Commission on the Future of Assessment in Education. (2013). *To Assess, To Teach, To Learn: A Vision for the Future of Assessment in Education.*

Gordon, E. W., & Bonilla-Bowman, C. (1996). *Can performance-based assessments contribute to the achievement of educational equity?* In J. Baron & D. P. Wolf (Eds.), Performance-based student assessment: Challenges and possibilities. Ninety-fifth Yearbook of the National Society for the Study of Education. University of Chicago Press.

Hanh, T. N. (2007). *Teachings on Love.* Parallax Press.

Jones, E. (1994). Emergent curriculum. *National Association for the Education of Young Children.*

Koretz, D., et al. (1993). *The reliability of the Vermont portfolio scores in the 1992–1993 school years* (Interim Report). CSE Technical Report. Rand Institute.

Ladson-Billings (1994). *The Dream Keepers: Successful Teachers of African American Children.* Jossey-Bass.

McAvinia, C. (2016). *Activity Theory.* In: Online Learning and its Users. Elsevier, pp. 59–100.

Moll, L. C., & Greenberg, J. B. (1992). Creating zones of possibilities: Combining social contexts for instruction. *Vygotsky and education: Instructional implications and applications of sociohistorical psychology*, *319.*

Pressley, M., & Ghatala, E. S. (1990). Self-Regulated Learning: Monitoring Learning from Text. *Educational Psychologist, 25*(1), 19–33.

Resnick, L. B., Asterhan, C. S. C., & Clarke, S. N. (2018). *Accountable talk: Instructional dialogue that builds the mind. Educational Practices Series, No. 29.* UNESCO International Bureau of Education/International Academy of Education.

Saturno, W., & Wolf, D. P. (1997). Archaeology and cultural exploration. *Digging Deep: Teaching Social Studies Through the Study of Archaeology. Portsmouth, Heinemann*, 1–23.

Schwebel, M., & Raph, J. (1973). *Piaget in the classroom.* NYC: Basic Books.

Vargas Llosa, M. (1991). *A Writer's Reality.* United Kingdom: Faber & Faber.

Vygotsky, L. (1962). *Thought and language.* (E. Hanfmann & G. Vakar, Eds.). MIT Press.

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky, Volume 1: Problems of general psychology* (pp. 39–285). Plenum Press.

Weber, J. (1946). *My country school diary.* Dell, New York, NY.

Wiggins, G. (2005). *Understanding by design.* Grant Wiggins and Jay McTighe. Expanded 2nd edition. *Association for Supervision and Curriculum Development.* Pearson.

Removing the "Psycho" from Education Metrics

Stephen G. Sireci and Neal Kingston

Abstract

In this chapter, the authors provide a brief overview of the traditional psychometric concepts of scaling and calibration and then describe more contemporary notions that leverage technology to (a) develop complexity scales for items and tasks, and (b) provide actionable information for teachers and learners. Different methods of reporting assessment results, and how those results can support learning, are also discussed.

The field of "*psychometrics*" has been described as "the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality traits, skills, and knowledge" (American Psychological Association et al., 1985, p. 93). Dissecting the word into its constituent roots, "psycho" refers to the measurement of unobservable characteristics of people, and "metrics" refers to measurement. The process of measurement requires a scale, so what is measured can be quantified. Unfortunately, measuring unobservable attributes is difficult, and proper quantification of those attributes is even more complex. The approximately 150-year history of psychometrics has reflected that complexity, and in doing so may have prevented educational assessments from reaching their full potential to help students learn. How can educational assessments help students learn? We believe one way is by providing clear information about student learning that can help both learners and their teachers know what students know and what to do next. However, providing clear and actionable information from educational tests has not been a strength of psychometrics.

In this chapter, we review some of the history and terminology of psychometrics relevant to educational assessment and illustrate new ways of using assessment information to better serve learners. Some of the new directions we suggest involve removing the layer of complexity that has traditionally come with scaling educational assessments. We posit that the primary purpose of educational achievement testing should be to enhance student learning by providing valuable feedback that can guide both students and teachers. When tests are designed to identify areas of strength and weakness, they help students understand what they have mastered and where they need to focus their efforts. This targeted feedback allows students to take ownership of their learning and make informed decisions about how to improve. For educators, test results can highlight which concepts need to be revisited or taught differently, ensuring instruction is responsive to student needs.

Properly designed educational assessments can foster educational achievement by emphasizing progress and improvement over time. Instead of viewing tests as immutable arbiters of ability, students can see them as opportunities to demonstrate growth and learn from their mistakes. This shift in perspective can reduce test anxiety and encourage a more positive attitude towards learning. By focusing on the *formative* aspects of testing, educators can create a supportive environment where students feel motivated to engage deeply with the material and strive for continuous improvement.

Moreover, when the primary goal of testing is to enhance learning, it aligns assessment practices with the broader educational mission of developing lifelong learners. Tests that are integrated into the learning process, rather than being isolated events, can help students develop critical thinking skills, problem-solving abilities, and a deeper understanding of the subject matter. This approach ensures testing is not just a measure of what students know at a single point in time, but a tool that actively contributes to their ongoing educational journey.

To properly promote new ways of using assessment results to support learning, we first briefly describe the current influence of psychometrics in educational assessment. Thus, we begin with some brief history to illustrate why psychometric perspectives have dominated the educational assessment field. We then introduce some key terms used in reporting the results of educational assessments. The final sections of the chapter describe what we believe to be particularly effective ways to report the results of educational assessments to help learners learn.

Relevance of Psychometrics to Educational Assessment

Why are most, if not all, educational and psychological test results reported on score scales? To answer this question, we must go back to the first experimental psychology laboratory established by Wilhelm Wundt at the University of Leipzig (Germany) in the 1880s. Working in Wundt's lab was Ernst Weber, who measured people's perceptions of physical stimuli such as weight, temperature, and pressure under strictly controlled conditions. He found when physical stimuli were increased or decreased (e.g., when additional weight was added to a scale) participants in his experiments did not always notice the increase or decrease until some threshold (of increase or decrease) was reached. He called this threshold the "just noticeable difference," and modeled the relationship between "stimulus intensity" and "perceived intensity" using a simple ratio. Fechner was highly influenced by this work because he was convinced the mind existed independently of the body and was looking for a way to prove it. He extended Weber's work using a logarithmic formula with the unit of measurement being these "just noticeable differences." We will skip the details here (which can be found in most introductory psychology textbooks or in Sireci et al., 1998), but what is important to note here is Fechner used these "just noticeable differences" to develop the first "psychological scale." It was this scaling process that laid the claim for psychology as a legitimate, quantifiable science.

Thus, the beginning of psychological measurement required a scale to quantify what was being measured. As the science of educational measurement evolved, this notion of scaling was adopted. The earliest forms of large-scale educational tests ordered people along the score scale based on their test performance. These scales were designed to capture "individual differences," and given the strong influence of Charles Darwin's work on variation and genetic evolution around this same time, much of the focus in early psychometrics was using measurement to understand the psychology of individual (and group) differences, rather than on understanding *each* individual. Francis Galton, Darwin's cousin, led that effort, which eventually became known as the eugenics movement (Sireci & Randall, 2021).

Today, the idea of using intelligence tests to categorize and order groups of people has been resoundingly refuted, in large part because the creation of such tests is culturally dependent (Malda et al., 2010). However, the idea of placing people on a continuous score scale based on a test endures via a scaling process called *item response theory* (IRT). In the next section we briefly explain traditional and IRT scales, which are the means for describing learners' performance on educational tests.

What is a Test Score Scale?

There are many different ways students' performance on an educational test can be reported. In most cases, the communication of a student's performance on a test will be reported on some type of numerical scale. In this section, we provide a conceptual overview of traditional test score reporting scales, as well as their strengths and limitations.

How are the Results of Educational Assessments Typically Reported?

The results from educational assessments can be reported in many ways (See for example, Linn & Gronlund, 2005; Zenisky & Hambleton, 2016). The simplest way to report test takers' performance on a test is to provide the number of points earned. For example, if a test is worth 50 points, and a student earned 30 points, the student's test score may be reported as 30. This simple score is often referred to as the *raw score* because there is no transformation of this simple count of the number of points earned. Thus, raw scores are not on a scale per se, and are determined solely by the number of items on test, the number of points each item is worth, and the number of points earned by the student.

Taking into account the maximum number of points that could be earned is often done by reporting the *percent correct score*, which is the percentage of points earned by the test taker divided by the maximum number of points that could possibly be earned. For example, 30 points out of a maximum of 50 points would be reported as a percent correct score of 60%. By accounting for the maximum number of possible points that could be earned, the percent correct score is an intuitive scale that ranges from 0 to 100—the percent correct score scale.

Raw scores and percent correct scores are often criticized for having little meaning because their interpretation is completely bound to the specific set of items on a test. For this reason, raw scores and percent correct scores cannot be compared across different tests, even if the different tests are designed to measure the same knowledge or skills. For example, "90% correct" may seem like a good score on a difficult test, but not such a good score on an easy test. Thus, raw score and percent correct metrics fall short for many testing purposes, such as tracking learners' performance over time, or using different sets of items (test questions) to avoid mistaking memorization for learning. To address this problem, *scale scores* are used on tests designed for more enduring purposes such as monitoring performance over time or generalizing learners' performance to a wider domain of knowledge and skills. Scale scores are test results that are reported on a "common" scale, so when different students take different tests, their performance can be meaningfully compared on the same scale.

The process of creating a score scale for educational tests uses the statistical concept of *deviation*, which reflects differences, or distances, of test takers from a focal point, such as a mean test score, or from one another. In educational assessment, this scaling perspective is known as *norm-referenced* testing, and must be distinguished from *criterion-referenced* testing, which is another way to report the results from educational assessments. Thus, before describing how score scales from educational assessments are typically created, we first describe the difference between these two assessment orientations. Understanding the perspectives of and differences between norm- and criterion-referenced testing is perhaps the most important knowledge needed to properly interpret, understand, and explain learners' performance on educational tests.

Norm-referenced and Criterion-referenced Interpretations of Test Performance

In the 1980s, one of the most popular shows on television was *Cheers*. The setting for the show was a neighborhood bar in Boston called, you guessed it—Cheers. What does this show and this bar have to do with interpreting test scores? There were two diehard customers who were always in the bar—Norm and Cliff[1]. They were the "average Joes" who hung out of the bar, talked to everyone, were nice to everyone, nosey, and so forth. Although they argued with each other they were friends. Norm has the perfect name because he was average at best—average meaning he was a "normal" guy. In the same way, norm-referenced score interpretations are referenced to the average—to the Norm. In statistics or sampling, the norm can refer to a standard normal distribution or some expected value of average. Thus, one way students' test scores can be interpreted is how far they are from this average—from this norm.

Norm-referenced testing grounds interpretation of learners' performance with respect to a *norm group*, which is a sample that sufficiently represents the population to be tested. Using this representative sample, learners' performance on a test is interpreted by how well they did relative to the norm (e.g., did she score above or below the mean?), or against specific groups of people (e.g., did she score in the top 10% of all students who took the test?) The norm group is typically established at a particular point in time. That is, the nationally representative sample tested in the norming year serves as the reference group until the test is re-normed. Percentile ranks and similar derivations of raw scores can be used to report norm-referenced information.

The normal curve is typically used to derive scale scores on educational tests. A *standard normal deviate* scale converts raw scores from a norm group to deviation scores, based on how far a learner's score is from the mean of the norm group, in standard deviation units. The formula for this transformation is,

$$SS = \frac{X-\mu}{\sigma} \quad [1]$$

where SS is the scale score that corresponds to the raw score X, μ is the mean raw score of the norm group, and σ is the standard deviation of raw scores in the norm group. On this scale, a score of -1 indicates a raw score that is one standard

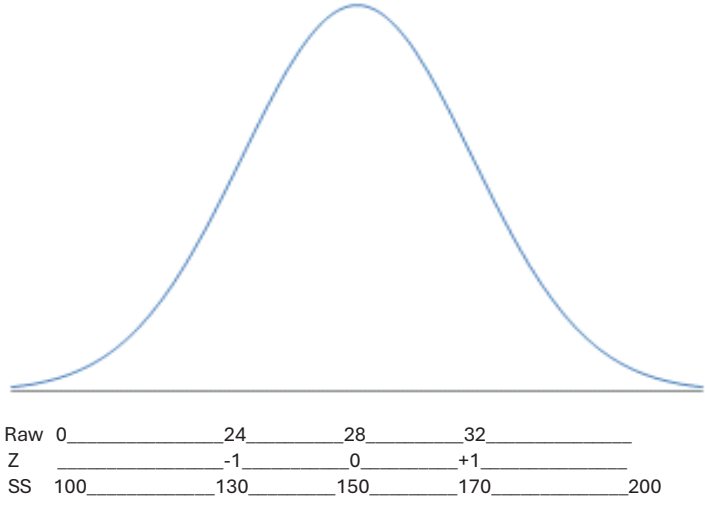
1 Not to be confused with Norman Cliff who wrote an excellent multivariate statistics book in 1987.

deviation below the norm group, and a score of 1 indicates a raw score that is one standard deviation above the norm group. Most score scales transform this normal deviate scale to one that avoids negative numbers and conforms to a scale considered to be more intuitive for interpretation. For example, the math section of the SAT college admissions test ranges from 200 to 800 with a mean of 500 and a standard deviation of 100 (based on a nationally representative sample of high school seniors in 2016—the mean and standard deviation may shift over time). Learners who score 650 on that scale have performed 1.5 standard deviations above the mean of the norm group.

An example of norm-referenced scales derived from the norming of raw scores is presented in Figure 1. This is a fictitious example from a nationally normed geography test on which there were 36 questions on the test form that was normed, and each question was worth one point. Thus, the raw score scale ranged from 0 (no items correct) to 36 (all items correct). Very few students did extremely poorly or extremely well on the assessment, which gave a rough normal distribution to the data. The mean of the norm group on this test was 28 and the standard deviation was 4. Under the raw score scale is the standard normal deviate scale (Z), and under that scale is a transformation of the normal deviate scale to report "geography test scale scores" (SS) that range from 100 to 200 with a mean of 150 and standard deviation of 20[2]

2 Transformations such as this one can be made using the equation of a straight line (i.e., a linear transformation), where SS=$bX+a$, and b the slope of the line) is equal to the ratio of the desired standard deviation (100) to the observed standard deviation (4), and a=$\mu-(b\bar{X})$. In this example $ss = [\frac{100}{1}x] + [150 - [(\frac{100}{1})(0)]]$, given the mean and standard deviation of the standard normal deviates are 0 and 1, respectively. This example illustrates how choice of scale is an arbitrary decision that can be used to report scores in a metric considered most acceptable by the testing agency.

Figure 1.
Illustration of Scale Score Transformations



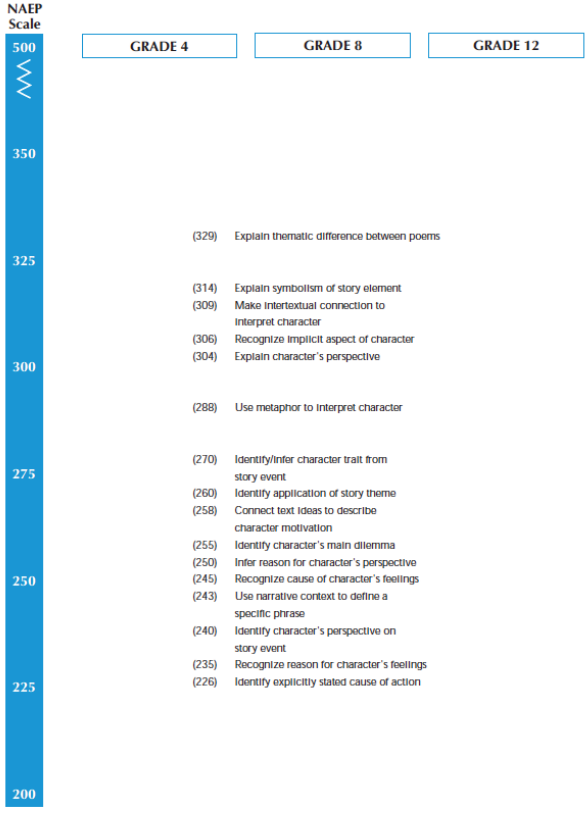
It should be noted the distribution of raw scores does not need to have a normal distribution for this transformation, and there is often an equating step involved when there is more than one test form involved. Those details are beyond the scope of this chapter and so readers who are interested in learning more about the psychometric activities of scaling and equating are referred to Angoff (1984) or Kolen (2006).

Item response theory (IRT) is an advanced probabilistic model for scaling educational tests that relates the probability of a learner correctly answering an item to the learner's location on the theoretical scale being measured. IRT also involves calibrating items onto the scale, which allows for learners to be placed on the same scale, even when they respond to different items. Readers interested in the details of IRT can refer to Hambleton, Swaminathan, and Rogers (1991) or Thissen and Steinberg (2020). We classify this method of scoring learners and placing them on a scale as norm-referenced, because to define the IRT scale, either the mean of the test taker population, or the mean of the item difficulties,

is arbitrarily set to a specific value, such as zero. However, because IRT locates each test item on the same scale as the test takers, actual items can be used to illustrate the knowledge and skills measured at specific points along the scale. This way of incorporating meaning into the score scale is referred to as *item mapping*. An illustration of item mapping is provided in Figure 2 (from Forsyth, 1998). This item map was used to help explain students' performance on the National Assessment of Educational Progress (NAEP) Reading assessment. Using the content standards measured by the items and the items' locations on the IRT scale, descriptions of the knowledge and skills students have at different points along the scale can be communicated. This procedure is suitable for large-scale tests because of the statistical requirements for IRT scaling. Where applicable, it can be used to understand skills students have mastered, skills they need to work on to achieve mastery, and subsequent skills they should work on next. Thus, although it is primarily a norm-referenced scaling procedure, IRT can be used to provide criterion-referenced information. IRT is also the underpinning of other scaling models, discussed in a later section of this chapter, that focus on providing criterion-referenced information.

Criterion-referenced testing does not involve relating learners' performance on a test to a mean or any other aspect of the population of learners. Instead, the information reported about performance on a test is referenced to the knowledge and skill domain targeted by the test. Criterion-referenced interpretations of test performance include statements like "Carlos mastered 82%" of the material on manipulating fractions" or "Yue achieved "proficient" status on the Grade 4 English Language Arts test." Unlike norm-referenced testing, in criterion-referenced testing, how other people performed on the test is not relevant to the interpretation of any one person's performance.

Figure 2.
Illustration of Item Mapping for Interpreting Students' Performance (from the National Assessment of Educational Progress Reading Test)



Source: Forsyth (1998).

Today, many criterion-referenced tests report learners' results using *achievement levels*. For example, the statewide summative (i.e., end of year summary) tests in Maryland classify students into one of the following four performance levels: beginning learner, developing learner, proficient learner, or distinguished learner. The "proficient learner" level represents what the state considers to be proficient for that grade level[3]. Theoretically, all students could be placed in the proficient, or any other, level. Other examples of achievement levels are "basic," "proficient," and "advanced," which are used on the NAEP (also known as the Nation's Report Card; see Loomis & Bourque, 2001).

The process of reporting learners' results in terms of achievement levels involves determining a level of performance on the test that is associated with each level. This process is described as setting "cut-scores" on the test or "standard-setting"; the latter term referring to the establishment of performance standards on the test associated with each achievement level (See Cizek & Ernest, 2016) for information on the process of standard-setting). The distinguishing feature of criterion-referenced information is it references learners' performance to a well-defined knowledge and skill domain, rather than to each other. Linn and Gronlund (2005) provide several guidelines for facilitating criterion-referenced information from educational tests. These guidelines focus on clearly defining the objectives of instruction and assessment, and ensuring the assessment provides sufficient information for determining whether students have mastered the objectives. They also suggest using item formats other than selected-response item formats (e.g., multiple-choice items) because students may correctly answer items by guessing.

It should be noted that tests designed to provide criterion-referenced information can also provide norm-referenced information. For example, in addition to reporting an achievement level for each student in Maryland, scale scores are also reported, and the mean performance of students in the school, district, and state is provided for parents, teachers, and others, to compare students' performance to these local and state averages.

3 The No Child Left Behind legislation and its extension the Every Student Succeeds Act requires all states receiving Federal funding to test students in reading, math, and science in several grades and to establish at least three achievement levels on each test, one of which must be "proficient" in that grade level.

The information provided by a test should match the intended purpose of the test. Therefore, the degree to which norm-referenced information from a criterion-referenced test is useful or holds substantive meaning would need to be established via research. There are also dangers, or potential negative effects, of reporting both norm-referenced and criterion-referenced types of information from educational tests. Norm-referenced information can reduce the academic self-concept of students who perform relatively worse than their peers, and some states use deficit-language in assigning achievement labels to students. For example, O'Donnell and Sireci (2021) reviewed all 50 states statewide summative testing score reporting practices and found that labels for the (same) lowest-performing level ranged from "inadequate" to "beginning learner." We were shocked to learn children were receiving score reports with such derogatory descriptions as "inadequate" and we support the work of Maryland and other states to promote more asset-based score reporting practices, which is a topic we return to in a subsequent section.

Table 1 summarizes the more traditional reporting practices for current educational tests. This summary provides a helpful baseline for us to compare the more innovative practices that are specifically designed to use educational assessments to promote learning.

Table 1.
Summary of Traditional Score Reporting Metrics for Educational Tests

Information Type	Reporting Metric	Explanation and Examples
Norm-referenced	Scale score	Distance of a student's score from norm group mean group in "standard deviation units" (e.g., Stanines (1–9 scale; mean=5); T-Scores (20–80 scale; mean=50)
	Percentiles	Percentage of students in norm group who score at or below student's score
	Age equivalent	Age at which the test score is "average" (e.g., 8.1= test performance is that of the average for a student who is a little over 8 years old).
	Grade equivalent (GE)	Grade in which the test score is "average" (e.g., a GE score of 4.5 is the average score of a 4th grader in the 5th month of school).
Criterion-referenced	Percent correct	Total points earned on a test divided by max possible points (e.g., 5-item test, each item worth 1 point, 4 items correct = 80%).
	Mastery score	A specific score signifying a student is proficient in the material tested.
	Achievement levels (Performance Classifications)	Specific levels of performance that describe various levels of mastery/proficiency (e.g., Pass/Fail; Basic, Proficient, Advanced)
	Subscores	Raw, percent correct, or scale scores for specific skill or content areas (e.g., persuasive writing; computations, applications, etc.)
	Item performance	Presents the items students took, the students' response, and the answer/scoring rubric.
Other	Gain scores	Difference between test scores taken at different time points (e.g., Spring score–fall score)

New Ways of Reporting Assessment Results to Help Learners Learn

The widespread use of commercial educational assessments that report a single score has made it clear that reporting a single score is ineffective in supporting student learning. With a single overall test score the actions an educator can take are limited and not particularly effective. For example, if a student has a low math score, a teacher can (a) continue to teach to the middle of the class and watch that student fall further and further behind, (b) provide that student with extra help, but with no knowledge of what lessons need to be reviewed, or (c) have the student repeat the grade. But as stated by mystery author Rita Mae Brown (1983, p. 68), though usually misattributed to Albert Einstein, "Insanity is doing the same thing over and over again and expecting different results." If a teacher does not know what it is that the student does not know, how can they efficiently and effectively remediate? Too often the primary outcome is disengaging students and reinforcing their belief that they cannot learn. Thus, there is a clear call for providing more instructionally valuable information from educational assessments.

First Came Subscores

The initial response to the need for more actionable educational test scores was subscores: separating out sets of items based on a test's content specifications and creating scores based on only those items (See Table 1). This separation can be done in several ways. For example, the ACT provides a composite scaled score based on all test items, but also provides math, science, English, and reading scores based on non-overlapping subsets of items as well as STEM and ELA scores that are based on combined item subsets.

One problem with subscores is they are based on a smaller number of items than total scores, and thus are less *reliable* (reliability refers to the consistency of test results). Another problem is they tend to correlate highly with each other, making it difficult to interpret score profiles (i.e., to distinguish between the different types of information provided). This correlation is magnified should some of the same items appear in multiple subscores. To minimize overinterpretation, some standards-based testing programs collapse score scale results to categories such as "below standard," "near standard," and "above standard." As discussed earlier another similar approach is to report subscores as either illustrating mastery or not mastery of the materials.

Backward Design to the Rescue

As an alternative to building a test and then deconstructing it to try to produce actionable subscores, *one could design a test with instruction in mind*. This alternative might be done by first designing the score report considered most likely to help teachers better teach students, and only then designing the test that would support that score report. Given that a goal of assessment in the service of learning is to help students learn and teachers teach, then perhaps an overall scaled score is not necessary. Perhaps teachers just need to know what each student has and has not yet mastered. Perhaps there might be a psychometric approach designed to do just that–and there is. We describe that approach next.

Diagnostic Classification Modeling

What is now usually referred to as diagnostic classification models (DCMs), but has also been referred to as cognitive diagnostic assessment, has much of its roots in the mid-20th century with the development of latent class models (Lazarsfeld, 1950; Lazarfeld & Henry, 1968). These models were initially used to identify unobserved subgroups within a population based on observed data. The concept of latent traits, which are not directly measurable, but can be inferred from patterns of responses, laid the groundwork for DCMs. Tatsuoka (1983) and Embretson (1984) specifically looked at decomposing cognitive processes within an IRT model to provide diagnostic information, which led to modern DCM.

DCMs provide a detailed analysis of students' knowledge and skills. Unlike traditional assessments that yield a single overall score, DCMs offer a nuanced profile of a student's strengths and weaknesses across multiple attributes or skills. This multi-dimensional approach allows educators to understand not just whether a student can solve a problem, but also which specific skills they have mastered or need to improve. By categorizing students into mastery or non-mastery for each skill, DCMs provide valuable diagnostic feedback that can be used to tailor instruction and interventions to meet individual learning needs. Figure 4, later in this chapter, provides an example of a score report consistent with a DCM approach.

The application of DCMs is particularly beneficial in formative assessments, where the goal is to monitor student learning and provide ongoing feedback. Teachers can use detailed diagnostic information to adjust their teaching strategies and provide targeted support. In large-scale assessments, such as state or national exams, DCMs offer a comprehensive picture of student performance across different

regions or demographics, informing policy decisions and resource allocation. Additionally, DCMs enhance adaptive testing environments by ensuring that the adaptive algorithm considers multiple skills simultaneously, providing a more accurate measure of student abilities.

Technically, DCMs are based on the assumption that the underlying traits being measured are categorical rather than continuous. This means students are classified into categories such as "mastery" or "non-mastery" for each skill. DCM models use a *Q-matrix* to define the relationship between test items and the underlying skills or attributes. The Q-matrix is a binary matrix that specifies which skills are required to correctly answer each test item. This matrix is crucial for the accurate estimation of students' skill profiles. The estimation process involves complex algorithms that can handle the multivariate nature of the data, often employing techniques from Bayesian statistics to provide robust and reliable classifications (Rupp, Templin, & Hanson, 2010).

Despite their advantages, implementing DCMs can be complex, requiring sophisticated statistical techniques and software. Educators and administrators need training to interpret and use the results effectively. Ensuring the validity and reliability of DCM-based assessments is crucial, involving rigorous testing and validation processes (Thompson, Clark, & Nash, 2021). Moreover, DCMs require detailed data on student performance across multiple items and skills, which can be resource-intensive to collect and manage. Nevertheless, the detailed insights provided by DCMs enable more personalized and effective educational interventions, ultimately supporting better learning outcomes. By offering a comprehensive profile of student abilities, DCMs represent a significant advancement in educational assessment, moving beyond traditional scoring methods to support more informed and targeted educational practices.

DCMs can be based on many approaches that identify skills and their relationship to mastery classifications as long as the result is a Q-matrix that captures those relationships. Skills can be initially identified in a single list or a hierarchy. Another, more recent, approach to identifying skills can be described as *learning maps*, which we discuss next.

Learning maps

Learning maps are one form of an organized learning model (Kingston et al., 2022) They are visual representations that illustrate the relationships among various knowledge, skills, and understandings within a subject area. They comprise an directed acyclic graph consisting of interconnected nodes, each representing a specific concept or skill. These nodes are probabilistically linked with the probability of having mastered a successor node conditionally dependent on having mastered one or more precursor nodes. Thus, the pathways among nodes show how different pieces of knowledge are related and how they build upon each other. Learning maps help educators and learners see the broader context of what is being learned and identify the pathways through which any individual student can best progress in their understanding.

Figure 3 shows a version (the learning map is updated regularly) of the Dynamic Learning Maps Alternate Assessment (DLMAA) mathematics map, which at the time the map was captured consisted of 2,554 nodes and 5,605 connections ranging from developmental infancy (recognizes an object) to advanced high school mathematics. With a map this large it should not be surprising that only the overall structure is captured in the figure. Those interested in how the learning maps for the DLMAA were developed can find more information in Bechard, et al. (2019).

Figure 3.
DLMAA Mathematics Learning Map

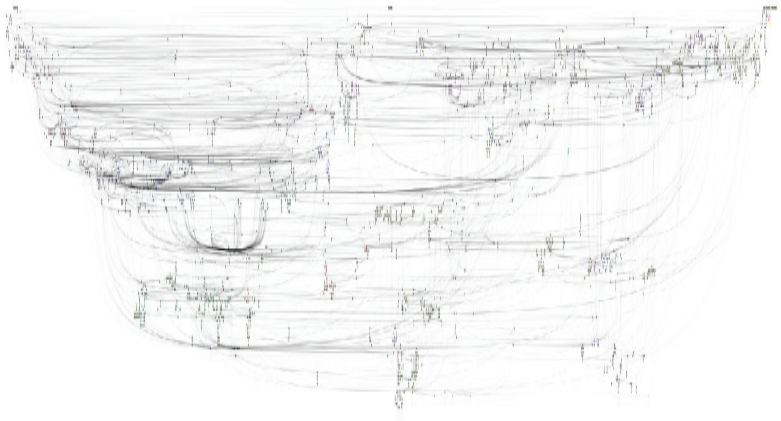
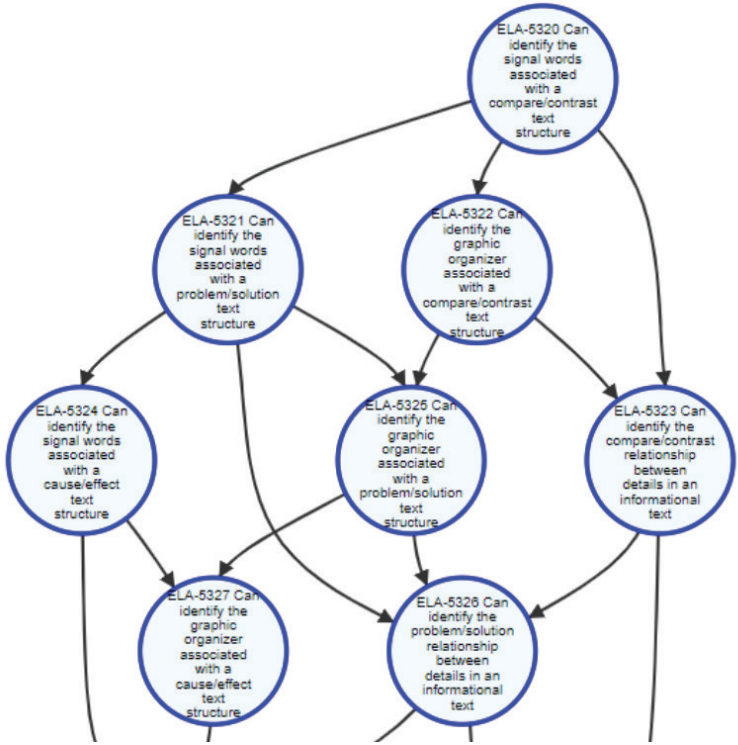


Figure 4 shows a section of an English Language Arts learning map that was developed as part of DLMAA. DLMAA will be described further in the next section of this chapter.

Figure 4.
Section of the DLMAA English Language Arts Learning Map



Because learning maps are directed acyclic graphs, they are amenable to diagnostic classification modeling including, but not limited to, an approach known as Bayesian network analysis. Bayesian networks are particularly useful for modeling complex systems where variables interact in uncertain ways, such as the human mind. They allow for efficient computation of the probabilities of

various outcomes given certain evidence, making them valuable for tasks such as diagnostic reasoning, prediction, and decision-making under uncertainty, such as which content to probe to best understand what a student knows and can do. Bayesian networks can incorporate both prior knowledge and new observed data to update inferences made from assessment data. In the next section, we turn to an example of learning maps applied to a specific situation—alternative assessments for cognitively impaired students.

Dynamic Learning Maps Alternate Assessment

The Dynamic Learning Maps Alternate Assessment (DLMAA) was designed to measure educational progress for students with significant cognitive disabilities. DLMAA was designed with different priorities than other accountability assessments. As stated by Kingston et al. (2014, p. 5), "For the world of educational assessment to better serve students with significant cognitive disabilities, we must begin with a goal for large-scale assessment that helps students learn." To accomplish this goal, six high-level features were chosen to guide the development of the DLM: "(1) fine-grained learning maps that guide instruction and assessment, (2) a subset of particularly important nodes that serve as content standards to provide an organizational structure for teachers, (3) instructionally embedded assessments that reinforce the primacy of instruction, (4) instructionally relevant testlets that model good instruction and reinforce learning, (5) accessibility by design (vs. accommodation) and alternate testlets, and (6) status and growth reporting that is readily actionable."

Learning maps play a crucial role in evaluating the educational progress of students with significant cognitive disabilities. The DLM system uses these maps to plot out individual concepts in nodes, showing the multiple ways that students' knowledge, skills, and understandings develop over time. This model helps educators uncover reasons why a student may be struggling with a particular concept and find possible pathways for students to expand their knowledge and skills.

The DLM assessments are designed to be adaptive and accessible, providing a more accurate measure of student proficiencies by adjusting the difficulty of questions based on the student's responses. The learning map model is integral to this process, as it ensures the assessment is aligned with the specific learning needs and proficiencies of each student. By mapping out the connections between different concepts, the DLM system helps educators identify the most effective

instructional strategies and provide targeted support to enhance student learning outcomes. This approach not only supports better academic performance but also helps in developing personalized educational plans that address the unique needs of each student.

DLMAA score reports

DLMAA reports scores by identifying precursor and successor nodes related to specific educational standards (essential elements in DLM parlance). These nodes are expressed as five levels in an extracted learning progression, illustrated in Figure 5.

Figure 5.
Section of DLMAA Learning Profile

Student's performance in 7th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during Spring 2022. Student was assessed on 13 out of 13 Essential Elements and 4 out of 4 Areas expected in 7th grade.

Demonstrating mastery of a Level during the assessment assumes mastery of all prior Levels in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

		Estimated Mastery Level				
Area	Essential Element	1	2	3	4 (Target)	5
ELA.C1.1	ELA.EE.RI.7.5	Understand the functions of objects	Identify concrete details in an informational text	Recognize how titles reflect text structure and text purpose	Understand sequencing	Understand how parts of the text affect overall text structure
ELA.C1.2	ELA.EE.RL.7.1	Differentiate between text and pictures	Identify characters, setting, and major events	Identify words that answer explicit questions	Identify where explicit information is stated and where inferences can be drawn	Identify explicit and implicit information
ELA.C1.2	ELA.EE.RL.7.4	Understand words for absent objects and people	Identify definition of words explicitly defined in a sentence	Identify word meaning of multiple-meaning words using context clues	Determine the meaning of idioms and figures of speech	Determine the connotative meaning of words and phrases
ELA.C1.2	ELA.EE.RI.7.2	Match a picture representation with a real object	Identify concrete details in an informational text	Identify the implicit main idea in an informational text	Identify multiple main ideas in an informational text	Summarize a familiar informative text

Levels mastered this year | No evidence of mastery on this Essential Element | Essential Element not tested

This report is intended to serve as one source of evidence in an instructional planning process. Results are based only on item responses from the end of year spring assessment. Because your child may demonstrate knowledge and skills differently across settings, the estimated mastery results shown here may not fully represent what your child knows and can do. For more information, including resources, please visit https://dynamiclearningmaps.org/states.

Page 3 of 5

The score reporting for DLM illustrates one current way for providing actionable information to help students learn. In the next and final section of our chapter, we discuss additional ideas for providing information to support student learning, as well as a research agenda for making progress toward this goal.

Criterion-Referenced Scales

In addition to using DCM to indicate the learning progressions of students, other research has focused on modeling the difficulty (challenge) presented by items and incorporating that understanding into the score scale. For example, Embretson (1993), Sheehan and Mislevy (1990), and other researchers (e.g., Fischer, 1973) have illustrated how the content, cognitive, and other "complexity" features of items can be used to understand and model the degree to which items challenge learners. These researchers have focused on relating item attributes, such as the number and type of calculations required to solve a math problem or the linguistic complexity of a reading passage, to the construct measured. In a sense, this work extends the item mapping described earlier (See Figure 2) to not only locate items on the scale, but to also quantify the specific knowledge and skills required to successfully answer items at different points along the scale. This research suggests the coding of content, cognitive, and other item features can be used to build a "complexity scale" to indicate where students are on the continuum of knowledge and skill measured by a test. The potential benefit of these complexity scales is the scale score reflects the complexity of the challenge presented by the items as they are located along the scale, rather than by how far students are from one another with respect to performance on a set of items. Quantifying score scales in this manner entirely removes norm-referencing from the scaling process and aligns students' performance with the cognitive challenges presented by the items (Feng et al., 2024). Future research is needed to evaluate whether reporting scores in this way will reveal more about the cognitive processes used by students in responding to questions and whether this information will support their learning.

Developing Educational Tests to Support Student Learning: The Path Forward

The inspiration and champion of the Handbook in which this chapter is written, Edmund W. Gordon, has long criticized educational tests as focusing on the status of learning rather than the process of learning. That is, educational tests tend to reflect *whether* students have learned, but have not been good at showing *how* they have learned. We believe the demonstration of performance along a criterion-referenced scale (e.g., item mapping) and subscore reporting based on learning progressions provide helpful information to support student learning, with the latter coming closer to the process of learning. However, to meet the challenge provided

by Gordon (2013, 2020), perhaps our assessing and reporting mechanisms should focus on representing processes, rather than only representing content.

As illustrated in Figure 4, the cognitive process of "identify" is a feature in the DCM for the learning map. Because DCMs map out a progression of skills, they have the potential to also map out processes, which could be a potential way to use assessments to provide information regarding the processes learners use to respond to questions, as well as the determining the likely success of different processes for individual learners. By understanding the processes learners used to answer questions, educators can "redirect" learners' to more effective processes or target the most important cognitive skills learners need to support their learning.

Another important area in need of further work is gathering empirical evidence on specific features of different approaches to assessment for learning, and the effectiveness of those features with the needs of different kinds of learners. Specifically, more research is needed on the best types of feedback to provide regarding student performance. In a meta-analysis of the effect of feedback on student learning (based on 435 studies of over 61,000 learners), Wisniewski, Zierer, and Hattie (2020) found a weighted mean effect size of 0.48, indicating a strong and positive effect of feedback. However, they also noted substantial variability due to context and type of feedback (e.g., reinforcement or punishment, corrective feedback, or high information feedback) and quality of feedback. Direction of feedback (teacher-to-student, student-to-teacher, or student-to-student) had only a moderate explanatory effect. Also, and not unexpectedly, there appears to be a publication bias in the published articles,[4] but none in the 116 dissertations. Dissertations showed a lower effect size of 0.36.

In short, while there is little question that overall feedback makes a positive difference, there is insufficient evidence regarding what *type* of feedback systems work best for which students. Various feedback models (e.g., Butler & Winne, 1995; Hattie & Timperley 2007) and their features need to be validated within educational assessment methods and systems. It is not enough to say tests should provide useful feedback; examples and strategies for providing useful feedback must be provided.

4 Peer-reviewed journals are notorious for their reluctance to publish non-significant findings, which biases the results of systematic reviews.

Another development in psychometrics that deserves more attention for targeting the processes learners use in responding to test items and using that information to serve learners is validity evidence based on response processes. The use of evidence regarding the cognitive processes students use to solve test items was largely championed by Messick (1989) and codified into the last two versions of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999; 2014). This source of validity evidence is used to support the use of a test for a particular purpose by probing whether learners use the intended cognitive processes when responding to items. Examples of such evidence come from focus groups, think-aloud studies, and cognitive interviews where test takers indicate the strategies they used in responding to items and report the thoughts and emotions they had while doing so (Padilla & Benítez, 2014). It would be interesting to not only use this information as validity evidence, but also as evidence to be reported to learners and educators as part of the results of an assessment.

Recently the analysis of "log data" from educational assessments has also been used to provide validity evidence based on response processes. Log data refers to the data a computer or other digital device captures as a test taker responds to assessment items. Like other methods for probing learners' response processes, these data are used in validity studies to confirm the intended cognitive processes are measured on an assessment, to check whether students are motivated to try their best when taking an assessment, and even as part of scoring an assessment. For example, He and von Davier (2016) analyzed log data from the computer-based Programme for International Assessment of Adult Competencies (PIAAC) assessment to understand how test takers solved problems on the assessment. Similarly, He at al. (2021) analyzed log data from a computerized test of problem solving and found it could be used to determine how far the observed processes used by test takers deviated from optimal problem-solving solutions. As another example, Wise et al. (2021) used the amount of time learners take to respond to test items as a measure of their engagement with the test. Chung et al. (this volume) provide other examples of how log data can be leveraged to understand more about the cognitive processes students use to solve test items. Clearly, these developments illustrate how log data from digital assessments can isolate and report the processes learners use. Access to this information can empower both learners and educators to adjust problem-solving behaviors in ways that maximize

success. Reporting information on engagement will also likely be useful for interpreting learners' test performance and for improving assessments that show low levels of engagement.

Another important area of the path forward is ensuring that the primary consumers of the results of educational tests—learners and their educators—clearly understand the results of an assessment. As O'Leary et al. (2017) pointed out,

> *score reports…are fundamental to the process of communication between test developers and their audience. As such, the interpretability of score reports (i.e., how well members of the intended audience are actually interpreting and using scores as reported) is of the utmost significance and is fundamental in claims about validity.* (p. 21)

Clearly, providing information on learners' performance on educational assessments must be done in a way that is understandable to learners and educators. In this chapter, we have argued that we must go beyond traditional score reporting practices to provide actionable information on both the status and process of learning. In keeping with the points of O'Leary et al. (2016), even the best intended extensions will be invalid if the audiences they are intended for cannot interpret them.

As a last suggestion for the path forward we provide an additional thought. If educational assessment in the service of learners requires that "Feedback, adaptation, and other relevant instruction should be linked to assessment experiences" (Principle 5), perhaps the results from educational tests should not so much focus on metrics, but focus on how they can best facilitate *conversations* between learners, teachers, and others. Perhaps even the results can facilitate conversations between learners and test developers—maybe even psychometricians! As described in Heritage and Kingston (2019) there has long been a divide between the approaches to assessment of classroom teachers and psychometricians. More and broader conversations are likely to lead to better tests that do even more to serve learners.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Educational Testing Service. (Reprint of chapter In R. L. Thorndike (Ed.). *Educational Measurement* (2nd Edition), American Council on Education, 1971).

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of Evidence-Centered Design to Develop Learning Maps-Based Assessments. *International Journal of Testing*, *19*(2), 188–205. https://doi.org/10.1080/15305058.2018.1543310

Brown, R. M. (1983). *Sudden Death.* Bantam Books.

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, T. Haladyna, & M. Raymond (Eds.), *Handbook of test development* (pp. 212–237). National Council on Measurement in Education.

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1984). *A general latent trait model for response processes. Psychometrika, 49,* 175–186.

Feng, W., Tran, P., McNichols, W., Sireci, S. G., & Lan, A. (2023, October). *Using artificial intelligence to scale multiple-choice mathematic items.* Paper presented at the annual meeting of the Northeastern Educational Research Association.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Forsyth, R. A. (1998). NAEP frameworks and achievement levels. In M. L. Bourque (Ed.), *Proceedings of the Achievement Levels Workshop.* National Assessment Governing Board.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145–221.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–112. https://doi.org/10.3102/003465430298487

He, Q., Borgonovi, F., & Paccagnellac, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*. 104170

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development*, (pp. 750–777). Information Science Reference.

Heritage, M., & Kingston, N. M. (2019). Classroom Assessment and Large-Scale Psychometrics: Shall the Twain Meet? (a conversation with Margaret Heritage and Neal Kingston). *Journal of Educational Measurement, 56*(4), 670–685.

Kingston, N. M., Alonzo, A. C., Long, H., & Swinburne Romine, R. (2022). Editorial: The use of organized learning models in assessment. *Frontiers in Education, 7*, 1009446. https://doi.org/10.3389/feduc.2022.1009446

Kingston, N. M; Clark, A. K., Pardos, Z., & Lee, S. Y. (April 2014). *Determining a Reasonable Starting Place for an Instructionally Embedded Dynamic Assessment: Heuristic versus Bayesian Network Analysis.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155–186), Praeger.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–472). Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Houghton Mifflin.

Linn, R. L., & Gronlund, N. E. (2005). *Measurement and assessment in teaching (9th edition).* Upper Saddle River, NJ: Pearson Prentice-Hall.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard-setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Erlbaum.

Malda, M., van de Vijver, F. J. R., & Tamane, Q. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38,* 582–595.

Messick, S. (1989b). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). American Council on Education.

O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment.* https://doi.org/10.1080/10627197.2021.2016388

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36(2),* 16–23.

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

Ryan, J. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T.M. Haladyna (Eds.), *Handbook of Testing* (pp. 677–710). Lawrence Erlbaum.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 255–272.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–135). Routledge

Sireci, S. G., Wainer, H., & Braun, H. (1998). Psychometrics, overview. In *Encyclopedia of biostatistics*. John Wiley & Sons.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation, 1,*(1), 23–39. https://doi.org/10.59863/GPML7603

Thompson, W. J., Clark, A. K., & Nash, B. (April 2021). *Technical Evidence for Diagnostic Assessments*. Paper presented at the annual meeting of the National Council for Measurement in Education (virtual conference).

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, and T. Haladyna (Eds.), *Handbook of test development* (2nd ed.), pp. 585–602. Routledge.

Using Learner-System Interactions as Evidence of Student Learning and Performance: Validity Issues, Examples, and Challenges

Gregory K. W. K. Chung, Tianying Feng, and Elizabeth J. K. H. Redman

Abstract

This chapter explores the idea of using learner-system interactions as a source of evidence about students' learning and performance in the context of Principle 3: Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition; and Principle 4: Assessments model the structure of expectations and desired learning over time. We illustrate how well-designed instructional opportunities in interactive digital environments naturally provide measurement opportunities. These opportunities can result in what we call "measurement without testing": Learner-system interactions that are designed to support students' learning are by definition observable and we believe carry the most relevant information about students' learning. Digital environments enable the collection of fine-grained behavioral data about what, when, and how a learner interacts within that environment.

However, for learner-system interactions to serve as evidence, three design challenges must be addressed: identifying the cognitive demands of the task, identifying the learning-relevant indicators of interest, and developing algorithms to transform low-level behavioral events into high-level indicators that represent learning-relevant processes. If we can observe what learners are doing as they do it and develop the methodology to accurately determine why, then that capability may help move us toward tailored, adaptive, and individualized learning for all students.

Author Note

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

Tianying Feng, ORCID: https://orcid.org/0000-0003-2215-9234

Elizabeth J. K. H. Redman, ORCID: https://orcid.org/0000-0002-5301-3716

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

The research reported in this chapter was supported by grants from the U.S. Department of Education's Ready to Learn program, the Institute of Education Sciences, and the National Science Foundation. However, research and findings do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the federal government. [PR/Award No. U295A150003, S368A150011, R305A190433, R305C080015, 2119818].

Digital environments enable the collection of fine-grained behavioral data about what, when, and how a learner interacts within that environment. The capability to automatically track the behavior of learners in digital environments has existed for years if the system was programmed to log such behaviors. The tracked behavior can range from learners' fine-grained, moment-to-moment behavior to the learners' final answer to a problem. In addition to behavior, the state of the environment can also be tracked and yoked to the learners' behavior.

The utility of tracking learners' responses has been recognized since the 1990s as a viable means to support the measurement of learners' processes and performance in interactive systems (e.g., Chung et al., 1999, 2002; O'Neil et al., 1997; Williams & Dodge, 1993; Young et al., 1997) using software sensors (Chung & Baker, 2003) and physical tasks using hardware sensors (e.g., Chung et al., 2021; Nagashima et al., 2009). Such data capture capability is routinely implemented in educational technology applications such as games, intelligent tutoring systems, training simulations, digital assessments, and large-scale standardized testing programs such as National Assessment of Educational Progress (NAEP) (Bennett et al., 2007; National Center for Education Statistics, 2012, 2020) and PISA (Foster & Piacentini, 2023; Organisation for Economic Cooperation and Development, 2014, 2021, 2023).

One of the most important reasons for tracking learners' behavior is to address questions related to how learners performed on a task, the processes they used to complete (or not) the task, and perhaps most importantly, why they performed the way they did (Feng & Cai, 2024; Jiao et al., 2021; Lindner & Greiff, 2023; Zumbo et al., 2023). Before we can address these questions, at least two conditions need to be satisfied: (a) availability of data on learners' responses in the interactive task *such that those data reflect learners' intentional behavior,* and (b) availability of information on the design features of the task, whether to promote learning or to test learners' knowledge or skills. While these two conditions are apparent for any assessment, it is less obvious how to satisfy them when the task is interactive and involves cognitive demands, including content knowledge, reasoning, and problem-solving processes.

Despite the long history and widespread use of digitally collected process data, there remain challenges in nearly every step of the analytics process: from specifying what behavior to record, how to capture it, the storage format, indicator specification, algorithm development, task design to support measurement, user interface design to support measurement, and incorporating theory into the entire endeavor. Lindner and Greiff (2023) outline key challenges and best practices for the use of process data for assessment purposes. They point out the need for a top-down approach to the design of assessments to ensure theory-grounded interpretation and analysis efficiency, and highlight the labor-intensive nature of process data analyses, including extensive data preparation.

In this chapter, we conceive of learner-system interactions—observable behavioral responses from the learner to some stimulus presented to them by the system, as well as the system's response to a learner's input—as the atomic unit of observation. In digital systems, this conceptualization flows from the capabilities enabled by software and hardware. Software or hardware can be developed to detect and log the learner's actions and system context (i.e., events and states) at the moment the action occurred and then save this packet of information to an external store.

Conceiving learners' interaction as an observation allows us to adopt well-established analytical frameworks and tools from measurement science. If the observation (or collection of observations) is used as a measurement, we can adopt a validity perspective to address issues related to the design and use of learner-system interactions and the design of tasks that can yield informative interactions. We use games as the specific context because of the complexity of interactions available in games (Chung, 2015; Chung & Feng, 2024; Lindner & Greiff, 2023).

A second reason to conceive learners' interactions as observations is that this conception directly addresses two of the Design Principles for Assessment in the Service of Learning:

- 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.

Conceiving learner-system interactions as observations focuses attention on the relation between the design of an interactive task and the learner's responses. Regardless of whether the measurement target is learning, motivation, attention, engagement, effort, or metacognition, how a learner responds to task demands is dependent on the degree to which the task is able to elicit a response representative of the target construct. Appropriate inferences of the learner-system interactions are dependent on the fidelity of the task. The key leverage afforded with learner-system interaction data is that fine-grained behavioral data are now available and using the observations as data should lead to a close examination of the alignment of the learner-system interactions, task design, and target construct—analogous to a test content analysis but at a much finer grain size (i.e., the level of the learner interacting with the digital task). As noted by S. Sireci (personal communication, December 20, 2024), similar to how close attention is given to how well a test represents the construct and how items are designed to measure the construct, learner-system interactions may be "another potential manifestation of the construct and the "new development" is how to capture the intended behaviors and ensure recording of the construct-relevant log data."

In the remainder of this chapter, we first define and present a detailed example of what we mean by learner-system interaction, demonstrating that even a simple game developed for preschool children has a rich set of interactions. We then discuss validity issues and underlying assumptions related to using learner-system interaction as an observation, highlighting the process of going from low-level clicks to an indicator. We then present a detailed example of the design process that led to a game design in which the game mechanics, originally designed to promote learning, could also serve a measurement function. Next, we discuss the challenges involved in using games for measurement purposes. We end the chapter with a brief discussion of outstanding issues and the relation of learner-system interactions to assessment in the service of learning.

Learner-System Interactions as the Atomic Unit of Observation

Modern digital systems are designed to attract and maintain users' attention, and interactivity is a key design feature. For example, learning games are highly interactive, making maintaining learners' engagement a critical design priority. With little engagement, there can be little learning, no matter the quality of the instructional material (Roberts et al., 2016). In contexts where users have a choice among different media, users will choose media designed to have more rather than fewer engagement elements (Roberts et al., 2016). An essential component of engagement is interactivity, which refers to the degree to which learner and system responses depend on each other (Domagk et al., 2010; Janlert & Stolterman, 2017; Kennedy, 2004; Plass et al., 2012).

Why Learner-System Interactions?

Using interactions as a potential source of evidence about learning is attractive for three reasons. First, as a practical matter, interactions can be captured via the software in digital systems. The software can be instrumented to log interactions. Well-designed instrumentation takes into account both the target cognitive demands and what the task allows learners to do (e.g., to engage in interactions that promote learning, to apply their prior or newly acquired knowledge, or to require reasoning or problem-solving to complete the task successfully).

The second reason for using interactions as a potential source of evidence is based on classroom interaction research, which shows robust findings that the nature of interactions between and among teachers and students can influence student learning (e.g., Greer & McDonough, 1999). Furthermore, a reciprocal relationship is established by the participants in the interaction, each being influenced by the other's action and the setting within which the interaction occurs (Young et al., 1997). Thus, how participants interact can determine what is learned (or not) and whether the interaction is productive (or not) (Young et al., 1997). The nature of the interaction—the extent to which a specific interaction episode is productively (or unproductively) related to the target outcome—helps explain why some students profit from instruction while others do not. A striking example is Webb's (1983) reanalysis of classroom interaction variables, which showed that examining only general interactions (e.g., giving or receiving help) led to no relation with achievement. However, when Webb recoded the interactions by type of help, the data revealed significant relations between the type of interaction and achievement.

The general finding that the quality of an interaction carries information about students' learning strongly suggests that learner-system interactions are promising sources of evidence of learners' knowledge and potential learning processes.

Third, the general methodology of behavioral observations has a long and robust research tradition. Interactions have been used as a data source in studies examining parent-child interactions, couples' communication patterns, teamwork processes, and classroom instruction (e.g., Bakeman & Gottman, 1997; Gottman & Notarius, 2000; Ostrov & Hart, 2013). Learner-system interactions are another form of behavioral observation using a technology-based collection of fine-grained behavior.

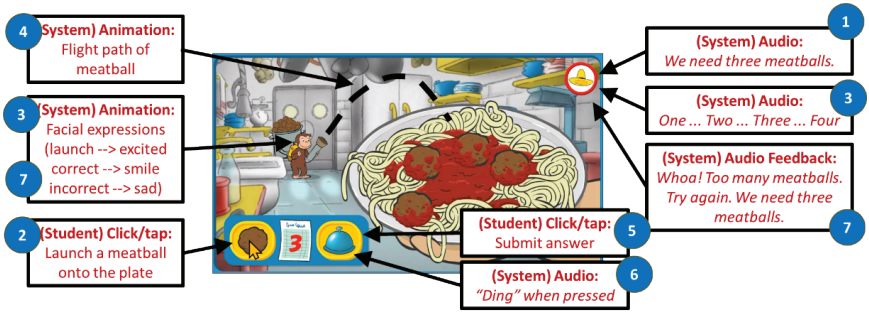
Example of Atomic-Level Interactions in a Learning Game

To provide a concrete example of what we mean by learner-system interaction, we present a simple illustrative example using *Meatball Launcher* (https://pbskids.org/curiousgeorge/busyday/meatballs/), a popular PBS KIDS game designed to expose preschool children to counting one to 5 objects upon request. The goal of *Meatball Launcher* is for players to add the number of meatballs specified by the target number shown on the screen to a plate of spaghetti. As shown in Figure 1, the level starts with a system voice-over giving directions (no. 1 in Figure 1). Players first click on the meatball (no. 2), and the system announces the current meatball count (no. 3). Curious George is the name of the monkey, and his facial expression changes from neutral to excited (no. 3, system animation), and he launches a meatball. The meatball flies from George to the plate (no. 4). The player can click on the meatball any number of times, even beyond the target number. When the player (presumably) thinks they have reached the target number, they click on the bell. The system responds with a "ding" (no. 6). The system then gives feedback to the player in two ways: a voice-over stating the attempt was correct or incorrect, and George's face (a smiling face for a correct solution and a sad face for an incorrect solution). The game automatically advances to the next level if the player is correct.

A close inspection of *Meatball Launcher* reveals the range of interactions in one level. For measurement purposes, the critical learning-system interactions are (a) clicking on the meatball button (no. 2 in Figure 1), (b) clicking on the bell (no. 5), and (c) solution correctness (no. 7). *Meatball Launcher* was instrumented from a measurement perspective (i.e., what game interactions could indicate players' counting skills?) and under the assumption that skill development could be

described by speed and accuracy. Thus, the following information was collected: timestamp of event, round number, target number, solution attempt, correctness of attempt, and text of the system feedback. From these data, we derived indicators of game progress as a proxy for speed (i.e., mean time per round and maximum round reached) and game performance as a proxy for accuracy (i.e., number of correct first attempts, number of correct attempts overall, and number of incorrect attempts).

Figure 1.
***System and Learner Interaction Elements in* Meatball Launcher**

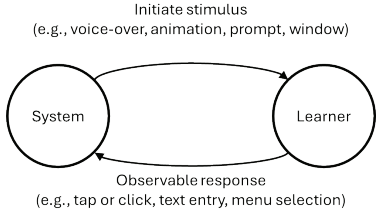
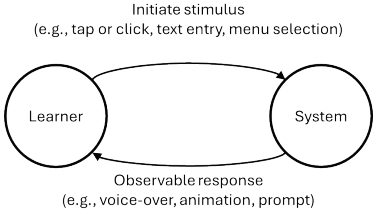


Note. https://pbskids.org/curiousgeorge/busyday/meatballs/

System-Initiated Interaction.

The left panel of Figure 2 shows a system-initiated interaction, where the system first presents some stimulus to the learner, and the learner responds to the stimulus through an action allowed by the user interface. The system-initiated interaction cycle represents a task design where the system needs input from the learner before the system can progress in the game, simulation, or assessment. The form of the learner's response is determined by the task design and expressed through a user-interface action.

Figure 2.
System and Learner-Initiated Interactions and Examples: Complete Interactions

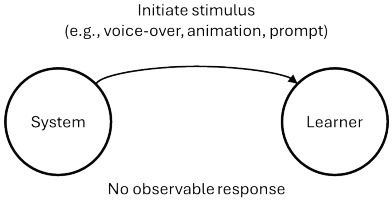
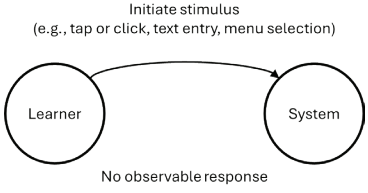
System-initiated Interaction	Learner-initiated Interaction
	
• *Meatball Launcher* sequence of events example • System audio (no. 1, directions) • Learner response (no. 2, clicks on meatball button) • System audio (no. 3, counts) • System animation (no. 3, facial expression) • System animation (no. 4, meatball flying) • Online multiple-choice example • The system presents the stem, options, and submit answer button • The learner responds by selecting an option	• *Meatball Launcher* sequence of events example • Learner responds (no. 5, clicks on the bell) • System audio (no. 6, ding) • System audio (no. 7, facial expression) • System audio (no. 7, feedback on correctness) • Online multiple-choice example • The learner clicks on the submit answer button • The system acknowledges the submission of the answer but does not provide correctness feedback

A typical system-initiated interaction is for the system to present a dialog or modal window. The window prompts the learner to make a decision or provides information, and the window cannot be dismissed without the requested action. In an online multiple-choice test, a stem is presented with a multiple-choice item, and the learner chooses an item option. The system can use the learner's inputs to determine the next item to show (in a computer-adapted test) or for feedback (e.g., to acknowledge acceptance of the answer submission). Similarly, in a game, the game may present a dialog for players to select a level to play. Note that the stimulus can be explicit or implicit and use audio, text, images, or graphics. Regardless of the media used, the underlying interaction is initiated with the system and ends with a learner response.

Learner-Initiated Interaction.
The right panel of Figure 2 shows a learner-initiated interaction, where the learner performs some action and the system responds to that action. The learner-initiated interaction cycle allows a task design to be open-ended and allows the learner to decide what action to take and when. The learner's input and the system's response formats are determined by the task design and expressed through the user interface.

The learner-initiated interaction is well-suited for open-ended tasks, where learners may have many potential actions—and thus decisions—to make. This type of task design is often used in digital performance tasks, games, and simulations. If the sequence of operations is important, then this task design can reveal the extent to which learners know or can determine the proper sequence. Likewise, if efficiency is important, then this task design may reveal economy of expression and differentiate between learners who know an existing solution to a problem from those who do not, and from learners who learn the solution over the course of the task. Finally, interactions may be incomplete where the system or learner initiates an action but no response is given as shown in Figure 3.

Figure 3.
System and Learner-Initiated Interactions and Examples: Incomplete Interactions

System-initiated Interaction	Learner-initiated Interaction
	
• *Meatball Launcher* sequence of events example • The three system-initiated events after the user clicks on the meatball button are examples of this type of interaction where no user inputs are expected. • Online multiple-choice example • Ancillary directions may be given as part of the task (e.g., check your work; keep track of remaining time) where no input is expected of learners.	• *Meatball Launcher* sequence of events example • The learner clicks on parts of the screen that are not designed to receive learner responses. Such off-clicks can be helpful when examining user-interface design (e.g., whether a button is too small or the hit point ambiguous). For example, learners unfamiliar with the interface may think clicking on George will initiate the launching of meatballs. • Online multiple-choice example • Many off-clicks may indicate learners are exploring the system, are bored, or want to exit the test.

Measurement Implications

Conceptualizing learners' behaviors in a digital system as interactions allows us to interpret behavior as a manifestation of cognition—one's choices in a task reflect one's knowledge and thinking. Because we can only observe learners' behavior and must infer the learning processes they use, tasks create situations for learners to demonstrate the use of the target cognitive demands. There needs to be user-interface elements that allow learners to interact with the system in a way consistent with the cognitive demands. For example, suppose a learning game is intended to promote problem-solving and we want to measure learners' problem-solving processes. In that case, the game should present situations where the learner is unlikely to immediately know the solution and provide information sources (e.g., resources, information, feedback, hints, and tutorials) that learners need access to and understand to solve the problem. To observe the problem-solving process, the information sources should be accessible via interactive user-interface elements instrumented to log the interactions. Learners will likely exhibit intentional behavior if the information sources are required to determine the solution to the problem. Problem-solving indicators can be derived from the learner-system interactions directly or through a transformation process.

The utility of learner-system interactions is threefold. First, viewing tasks as composed of learner-system interactions helps us describe general task features suitable for measuring different cognitive demands. Learner-initiated interactions (See Figure 2, right panel) may be well suited for assessing learning processes when the learner decides what to do next in a task. This design is akin to performance assessments. System-initiated interactions (See Figure 2, left panel) may be suitable when measuring specific knowledge or skills.

Second, learner-system interactions support quantitative analysis of the learner's performance and processes in a task. Some interactions may be directly evaluated (e.g., the learner's submission of an answer can be evaluated as correct or not in the example game *Meatball Launcher* as well as a multiple-choice task). Some interactions may need to be part of an algorithm that uses sets or sequences of interactions to derive an indicator, such as when examining learning over the course of the task. Regardless of the level of aggregation and transformations, the learner-system interaction is the atomic unit of observation.

Finally, digital systems directly record interactions whenever they occur, unlike traditional behavioral observations that typically use video or audio recording and rely on human coding of the data using a rubric. Data are generated each time learners perform an action. This data collection method can produce hundreds of interactions per learner regarding their game behavior. Even though these data "come for free," the situation creates a new set of validity concerns. In video coding, the transformation of events into a category or score is through the rater's interpretation of the scene relative to the rubric. The rubric can be inspected and critiqued in light of theory or construct. Inferencing is left to the human rater. In contrast, generating an indicator or score from interaction data is through the coding of algorithms. Data elements are extracted from the raw interactions, transformed, and eventually, a quantitative value is computed for a learning-related variable.

Validity Issues Related to Using Interactions as a Source of Evidence: From Clicks to Constructs

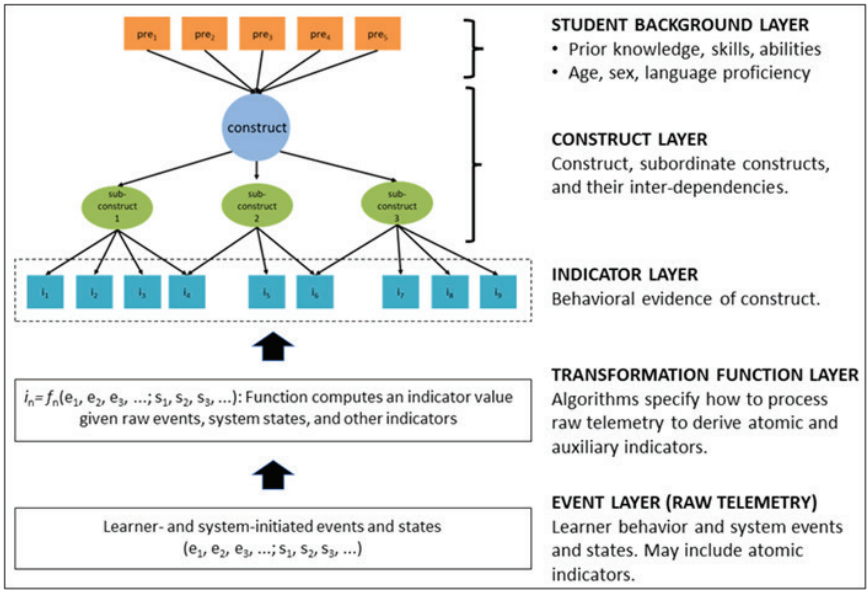
The process of transforming learner-system interactions into an indicator is shown in Figure 4 (Chung & Feng, 2024). In their discussion related to Figure 4, Chung and Feng expressed concern about the difficulty and amount of programming required to transform low-level interaction data into meaningful indicators. The authors asserted the (strong) assumptions encoded in the indicator development process. The assumptions were based on the validity issues identified by the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014), Baker (1997), and Mislevy et al. (2015) and applied to the situation of using learner-system interactions as evidence of learning. The following list of assumptions underlying the use of learner-system interactions for measurement purposes is from Chung and Feng (2024):

- The construct is an abstraction of human cognition and is not directly observable (AERA et al., 2014; Cronbach & Meehl, 1955; Messick, 1995).
- The construct has well-defined boundary conditions (AERA et al., 2014; Messick, 1995). For example, a clear definition describes the domain, the dependencies between and among the components of the domain, and an explicit relation between the construct (or subconstruct) and observable responses (Mislevy et al., 2015).

- The existence of the construct manifests in learners' generating observable responses. Responses include neurological, physiological, and motor responses at the lowest level; however, we are referring to the level of intentional behavior, speaking, or writing (e.g., see evidence-centered design, Mislevy et al., 2015).
- Learners are malleable (i.e., may learn) with respect to the construct and components of the construct, and their learning is influenced by what they observe, experience, perceive, or imagine. While the possible stimuli span the five senses, we are referring to visual, audio, and haptic inputs typical of learning settings, which may involve various types of static or interactive media, technology, real-world situations, or other people. By malleability, we mean, for example, that learners' skill in adding unit fractions can improve under certain conditions (e.g., when "good" instruction is provided and the learner is attentive to the instruction, exerts effort at processing the instruction and uses productive learning strategies).
- Learners' observable responses covary with changes in the construct in explainable and predictable ways. This assumption directly impacts measurement. If the learners' responses do not change even if they are learning, then it will be impossible to detect learning no matter how sensitive the measurement instrument is. If the learners' responses change for reasons unrelated to the construct, then the measurements will have little meaning. Finally, if learners respond unpredictably when the construct changes, the measurements will be unreliable and may indicate poor construct definition, poor choice of what is observed, or both.

Figure 4.

Conceptual Framework of the Relations Among Telemetry, Algorithms, and Indicators



Note. Telemetry is synonymous with learner-system interaction.

Event Layer

The lowest layer in Figure 4 is the event layer. The event layer comprises the learner-system interactions, the atomic unit of observations. The learner-system interactions are fine-grained data generated when a user behavior occurs. Note that the software must be instrumented to capture each learner-system interaction. Without instrumenting the software, no behaviors can be logged.

The event layer is important because it provides the raw data on which all other layers are built. The choice of what interaction to log affects what indicators can be derived, what analyses can be conducted, and ultimately, what inferences can be drawn about players. The key design guideline is to log learner and system

interactions representing learners' productive and unproductive use of the target knowledge or learning process. State information at the time of the event helps to disambiguate the action or aid in the subsequent creation of auxiliary indicators. See Chung (2015) for additional telemetry design guidelines.

Transformation Layer

The transformation layer in Figure 4 defines indicators in terms of algorithms. Given a definition, the algorithm derives indicator values from the data provided by the event layer.

The transformation layer is important because it provides inputs to a statistical model or procedure, from which inferences of learning are drawn. The transformation layer highlights that a processing stage is needed to transform raw interaction data into indicators—a stage that is often unreported, downplayed, or ignored in the literature. This processing stage is where coding and algorithm development occur. The specifications for the algorithms may be based on theory (i.e., hypothesized behavior under certain conditions), prior research that describes actual behavioral responses under certain conditions (e.g., see Feng's [2019] implementation of Metz's [1993] descriptions of misconceptions related to a pan balance), or data-driven approaches. The algorithm must be made available for inspection and critique because in this layer, solely behavioral responses (i.e., learner-system interactions) are transformed into indicators of learning processes and states that are otherwise unobservable.

The following section presents a detailed example of a game developed to promote learners' understanding of fractions. Baker's model-based assessment framework (Baker, 1997) was used as the general design approach, and Mislevy's evidence-centered design (Misley et al., 2015) was used to focus the linkage among observables, work products, and domain model.

Illustrative Example

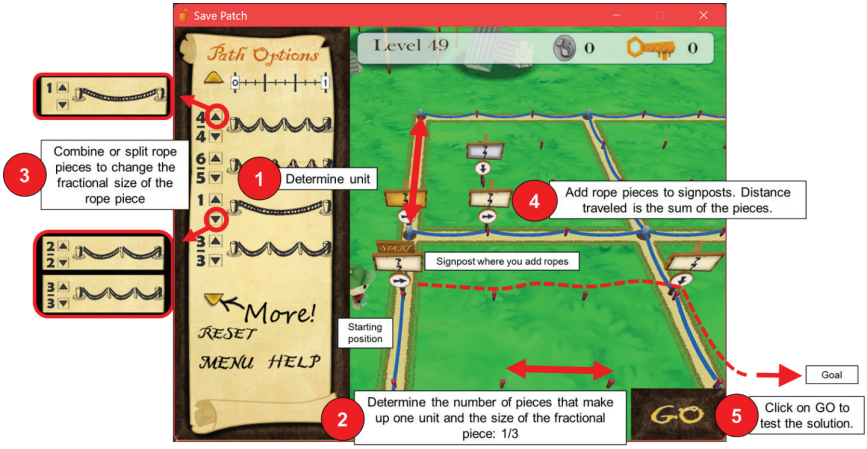
The example game *Save Patch* was developed by the Center for Advanced Technology in Schools (CATS) & CRESST (2012). *Save Patch* was designed to support middle-schoolers' learning of rational number equivalence (i.e., fractions). We assert that behaviors, expressed as learner-system interactions during the process of learning, can also be used for measurement purposes. The more the

instruction is aligned to explicit learning goals, the more information the interaction carries because the learner will be engaged in processes directly related to the target learning constructs.

Save Patch Game

Save Patch is an example of a game with game mechanics designed to address target learning outcomes directly. This example also shows how game mechanic interactions, originally designed to facilitate learning, can be used for measurement purposes. Based on the 2008 National Mathematics Advisory Panel (NMAP) report, UCLA/CRESST designed and developed a series of games (with *Save Patch* as the best example) to target two core ideas. The first idea is that all rational numbers (integers and fractions) are defined relative to a single unit quantity. The second idea is that rational numbers can be summed only if the unit quantities are identical (e.g., 1/4 + 3/4 is permissible, but 1/2 + 3/4 is not because the units or sizes of the fractions are unequal). Figure 5 shows a screenshot of the game.

Figure 5.
Screenshot of Save Patch Level 49 of 52



User Interface, Gameplay, and Learner-System Interactions. In *Save Patch*, the setting is an archeological dig site, and the player must help the game avatar retrieve a cat statue some distance away. The avatar can only travel along a one- or two-dimensional grid. The player lays out a path for the avatar by connecting signposts with ropes. The learner-system interactions include selecting the rope piece size and adding the correct number of rope pieces to a signpost. A submit button is included so the player can test the solution. The game only allowed rope pieces of the same size (denominators) to be added together. Fraction manipulation complexity was increased over levels through the grid spacing and rope sizes. For example, in more complex levels, players needed to subdivide two ropes until both ropes had pieces of the same fractional size (i.e., same denominator) (e.g., rope 1: split 1 into 6/6; and rope 2: split 1/2 into 3/6). Table 1 shows the relation between the fraction concepts, the game representation, and the associated learner-system interactions.

Table 1.

Relation Between Fractions Knowledge and Learner-System Interactions in Save Patch

Fractions concept [a]	Game representation	Cognitive demand and learner-system interaction
A unit can be represented as one whole interval on a number line.	The unit definition of the given grid is indicated by large gray posts (item 1 in Figure 4).	Cognitive demands: • Identify two large gray posts and understand that distance represents the unit (item 1 in Figure 4).
• The size of a fraction is relative to how a unit is defined. • The denominator of a fraction represents the number of identical fractional pieces in a unit.	• Fractional grid pieces between the large gray posts delimited by small posts [i.e., the denominator] (item 2 in Figure 5).	Cognitive demands: • Determine the size of the fractional piece between two small posts (item 2 in Figure 5). • Determine the size of the rope piece that represents the fractional piece between two small posts (item 3 in Figure 5). • Split or combine the rope to get the appropriate fractional piece size [i.e., the denominator] (item 3 in Figure 5). Learner-system interactions: • Click the up or down arrow (item 3 in Figure 5).

Table 1. (Continued)

Fractions concept [a]	Game representation	Cognitive demand and learner-system interaction
• The numerator of a fraction represents the number of identical parts that have been combined. • The units (or parts of units) must be identical to add quantities.	The avatar needs to travel from the starting point to the goal. The path to the goal is along the grid marked by signposts. The distance between 2 signposts is the amount of rope pieces needed. Placing the correct number of rope pieces of appropriate sizes between all signposts along the solution path beats the level.	Cognitive demands: • Determine the path from the starting point to the goal by identifying the signposts. • Determine the direction to travel between signposts. • Determine the number of rope pieces between signposts on the solution path. Learner-system interactions: • Learner: Drag a rope piece onto the signpost (item 4 in Figure 5). • System: Reject rope pieces with denominators different from the pieces on the signpost. • Learner: Click on GO to test the solution (item 5 in Figure 5). • System: After the player clicks GO, the avatar walks along the solution path, traveling the distance in the signpost. If the value is incorrect, the avatar will not reach or will overshoot the next signpost and fail. Level success is indicated with a message indicating completion.

Note. Additional information was logged with each interaction, including a timestamp and the state of the game (i.e., contextual information that includes current level, grid size, grid spacing, level solution set, and interactions-specific information such as correct or incorrect action). [a] CATS & CRESST (2013a).

Addressing Validity

Earlier, we asserted that a game designed for instructional purposes could also be used for measurement purposes. This section describes the elements that make that dual use possible. We briefly describe the design components and the resulting game and game mechanics.

Coherent Design Process.

Save Patch was part of a randomized-controlled trial to test the effectiveness of instructional games on students' understanding of rational numbers (See U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2015, for a WWC review of the study design). To ensure alignment among instruction, assessment, and professional development, we identified the critical knowledge in rational numbers. The knowledge was gathered from pre-algebra ontologies (Baker, 2012), Common Core Math Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), recommendations of the NMAP (2008), and practicing mathematics teachers. A set of knowledge specifications for rational number equivalence was developed from these sources, and the specifications were used to guide the design of the game instruction, fraction knowledge measure, and professional development. Figure 6 shows a snippet of the specifications.

Figure 6.

Excerpt of the Knowledge Specifications Used in the Design of Save Patch

CATS Knowledge and Item Specifications: Rational Number Equivalence

Rational Number Equivalence Knowledge Specifications		**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?*	
		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.		
	1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.).	A histogram of a certain quantity represented by discrete objects	Identity the unit that each single discrete object represents (e.g. each rose represents thousands of flowers sold on Valentine's Day).	A relationship between a real world measure and a scale model	Explain how what size of unit to use on the model to accurately represent the real world quantity (e.g. 1 inch equals 25 feet since the real world measure is 100 feet and the model can be up to 4 inches in length).
	1.3.0 In our number system, the unit can be represented as one whole interval on a number line.	A number line labeled with consecutive integers that may or may not include zero	Show the unit interval that fits with the given number line or accurately place another non-consecutive integer on the number line.	A number line that is labeled by skip units (2,4,6, etc.) or a line labeled by ½ units that may or may not include zero	Explain how to determine where other integer and rational values should be placed.

Note. CATS & CRESST (2013a).

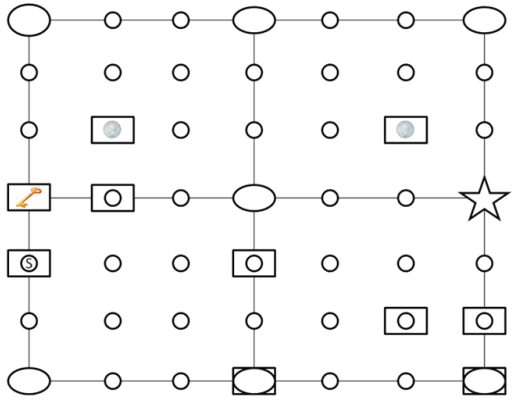
The game level progression was based on the mathematical development of fractions knowledge. The game followed a progression that introduced the game mechanics through tutorials and whole numbers. The game progressed from whole numbers to increasingly more challenging levels that involved complex fraction manipulations. Variation of practice was embedded by having multiple levels on the same topic (CATS & CRESST, 2013b). Table 2 shows the level sequencing. *Save Patch* had a total of 57 levels.

Table 2.
Level Sequencing for Save Patch

Math topic for each stage	Level number
Whole unit jumps; adding wholes	1
	Tutorial level on game mechanics
	2 to 4
Identifying correct denominator; scrolling to appropriate denominator, no adding fractions.	Tutorial level on fractions
	5
	Tutorial level on keys and coins
	6 to 9
Identifying correct denominator; inconsistent jumps (sometimes whole, sometimes fractional pieces)	10 to 13
Identifying correct denominator; jump over unit bar	14 to 16
Adding fractions; given correct size ropes	17 to 20
Adding fractions; jumps larger than one unit	21 to 24
Test levels: Add fractions when given different piece sizes.	25 to 28
Conversion of ropes; scroll given the wrong size	29 to 35
Conversion; Given smaller (e.g., 1/6) ropes to put on larger (e.g., 1/3) grid	36 to 38
Conversion; whole jumps, given fractional pieces of coils	39 to 42
Conversion; Common denominator needed	43 to 57

Finally, each level was documented to record the resources, grid layout, solution(s), and relevant knowledge specifications (CATS & CRESST, 2013b). Figure 7 shows the design for level 49. The representation describes the level, facilitating review and revisions during game development, algorithm development, and analysis phases.

Figure 7.
Game Level Design for Level 49



Level 49	Resources	Math Goals
- W x H = 2 x 2 - Grid unit = 1/3	- **6** 1/6-unit coils - **4** 1/4-unit coils - **11** 1/5-unit coils - **2** 1-unit coils - **3** 1/3-unit coils - **2** 1/2-unit coils - **10** directional signs	Conversion; Common denominator needed

Knowledge Specs Targeted in This Level

1.1.0. The size of a rational number is relative to how one whole unit is defined.
1.3.0 In our number system, the unit can be represented as one whole interval on a number line.
1.3.3. Positive non-integers are represented by fractional parts of the interval between whole numbers.
2.1.0. To add quantities, the units (or parts of units) must be identical.
2.1.2. Positive integers can be broken (decomposed) into parts that are each one unit in quantity. These single (identical) units can be added to create a single numerical sum.
2.1.3. Each whole unit or part of a whole unit (fractions) can be further broken into smaller, identical parts, if necessary.
2.2.0. Identical (common) units can be added to create a single numerical sum.
3.1.0. The denominator of a fraction represents the number of identical parts in one whole unit. That is, if we break the one whole unit into "x" pieces, each piece will be "1/x" of the one whole unit.
4.1.0. The numerator of a fraction represents the number of identical parts that have been combined. For example, ¾ means three pieces that are each ¼ of One Whole Unit.
5.1.0. The numerator is the top number in a fraction
5.2.0. The denominator is the bottom number in a fraction.

Game Solution Description

[(0/3,2/3), U, 4/12]; [(0/3,3/3), R, 4/12]; [(1/3,3/3), U, 4/12]; [(1/3,4/3), R, 16/12]; [(5/3,4/3), D, 12/12]; [(5/3,1/3), R, 4/12]; [(6/3,1/3), U, 8/12]

Game Mechanics That Closely Reflected Mathematical Operations.
Table 3 shows how the learner-system interactions described in Table 1 are transformed (or not) into indicators that can be used in an analysis procedure. Two indicators are shown: (a) game performance and progress indicators, and (b) fraction misconceptions.

The game performance and progress indicators are intended to be general indicators whose definitions can be adopted across different games and tasks. The game performance and progress indicators are intended to be general indicators, the definitions of which can be adopted across different games and tasks. Performance and progress are two common ways of describing human performance, from motor and verbal learning to outcomes, to education and training outcomes (e.g., Ackerman, 1990; Anderson, 1982; Fitts & Posner, 1967; Heitz, 2014). In games that have instructional sequences and learning-based interactions, we have consistently found these performance and progress measures to be associated with external criterion measures in expected ways (e.g., learners who know more, compared to learners who know less [as measured by an external criterion measure of knowledge or skill], demonstrate higher performance in the game, commit fewer errors, and spend less time on levels) (Chung & Feng, 2024).

We think these "common measures" are sensitive to knowledge outcomes because the game is intentionally designed to evoke learning processes. The learner-system interaction represents learners performing actions that use their existing or to-be-learned knowledge. The game levels are sequenced, where later levels build on what was learned in earlier levels. Games that do not require the learner to demonstrate the use of the target knowledge can yield interaction data, but the data will not likely reflect the use of the target knowledge. For games that lack a learning sequence (or curriculum), associating game progress with the degree of knowledge and skill will be tenuous; game progress may be a stronger indicator of engagement than learning.

The second kind of indicator in Table 3 is fraction misconceptions. These indicators show the utility of fine-grained interaction data but also highlight the challenge of using fine-grained data (discussed in the next section). See Chung (2015) for an example of how the data were structured. The learner-interaction data packet included as much information about the situation as we believed could be useful in as many analyses as possible. Given the knowledge specifications focused

on the whole unit and its composition of any number of equal-sized pieces (i.e., concepts of numerator and denominator) and that the addition operation could only be performed on pieces of the same size, we surmised it was important to record the specific numerator and denominator values along with the grid piece size for each addition operation. Further, while the correctness of the addition operation could give an overall indication of understanding, we reasoned that we would be able to identify particular misconceptions only with the exact numerator and exact denominator in relation to the specific level. Other examples that illustrate the use of fine-grained learner-system interaction data to infer cognitive processes are given in Chung and Feng (2024).

Table 3.

Selected Examples of Associations Between Learner-System Interactions and External Measure of Counting Knowledge (n = 783 to 851 middle school students)

Indicator	Validity evidence [a]	Learner-system interactions	Transformation
Game performance and progress indicators [b]			
Correct addition of fractions	0.10**	Add a rope piece to a signpost as often as needed to travel from one signpost to the next.	Because the game interaction closely resembled the act of adding two fractions, we evaluated the learner interactions as correct or incorrect. The indicator is the total number of correct or incorrect additions over the entire game. The game already computes correctness to determine whether the rope piece is permissible. Thus, correct or incorrect additions can be logged directly with no transformations.
Incorrect addition of fractions	– 0.19**		
Number of correct first attempts at solving a level	0.55**	Click on GO to test the solution Level success or failure	Because the game directly logged the results of the solution attempt, the only transformation was to filter the data for the first attempt of each level.
Number of correct attempts at solving a level	0.43**		No transformations were needed because the game directly logged the results of the solution attempt.
Number of incorrect attempts at solving a level	– 0.45**		

Table 3. (continued)

Indicator	Validity evidence [a]	Learner-system interactions	Transformation
Fractions misconceptions [c, d, e]			
Unitizing error		Add a rope piece to a signpost as often as needed to travel from one signpost to the next.	Because of the detailed logging of each operation (i.e., current level, grid size, grid spacing, level solution set, correct or incorrect action), unique tokens could be formed encoded the adding rope interaction event and the specific fraction values of the rope chosen by the learner. The tokens were then clustered using a fuzzy cluster algorithm. The clusters were labeled based on how the interactions comported with the extant research on fraction misconception.
Saw as one unit	– 0.28***		
Saw as wholes	– 0.22***		
Partitioning error			
Counted hash marks	– 0.21***		
Counted hash marks and posts	– 0.29***		
Unitizing and partitioning error			
Saw as one unit and counted hash marks	– 0.37***		
Saw as one unit and counted hash marks and posts	– 0.50***		
Iterating error			
Wrong numerator	– 0.44***		
Converting to wholes error			
Saw as a mixed number	– 0.11**		

*$p < .05$. **$p < .01$. ***$p < .001$.

[a] Spearman nonparametric correlation (ρ) between the indicator and an external measure of fractions knowledge. See Vendlinski et al. (2010) for a description of the measure. [b] Chung and Roberts (2018). [c] Chung and Feng (2024). [d] Kerr & Chung (2012). [e] Kerr (2014).

Challenges

In this section, we address what we see as three major challenges of developing indicators from learner-system interactions. While these challenges are discussed in relation to games, in our experience, the challenges surface whenever fine-grained data are used to infer high-level processes. Regardless of whether the collection system is software or hardware, or the task is game-based or not, we believe the challenges remain the same.

Challenge 1: Identifying the Cognitive Demands of the Game

Given a learning game, how do we examine the game and identify what the game is intended to teach? How do we determine whether a learning-system interaction (i.e., game mechanic) is useful for measurement? While game developers may be the obvious first choice, they are typically not trained in the learning or measurement sciences. The vocabulary used in the learning and measurement sciences may not mean the same thing to game designers as it does to learning and measurement specialists.

Addressing Challenge 1: Feature Analysis

Thus, one method to better understand the learning opportunities presented by a game is through an in-depth qualitative analysis ("feature analysis") of the game and its interaction opportunities. Feature analysis is the qualitative coding of an object (e.g., game, video, test item, intervention, assessment setting) against a set of properties. The properties are defined a priori (though often refined during the analysis process) and reflect aspects of the intervention hypothesized to influence student learning. The concept of feature analysis has its roots in Gordon (1970), in which he mentions qualitative analysis of assessments to describe cognitive functions to identify learning experiences required to promote positive academic outcomes more effectively. Subsequent development by Tatsuoka (1983) quantified this approach via her rule-space methodology, which mapped test items to a set of knowledge attributes to create a "Q-matrix." This item-attribute Q-matrix could then be subjected to quantitative analysis to examine, for example, the particular knowledge components (e.g., basic concepts and operations in fractions and decimals) (Tatsuoka et al., 2004). A key theoretical contribution of Tatsuoka's work was that the test items possessed certain attributes that could be reliably identified from a cognitive or knowledge perspective. A key CRESST insight was that this approach of describing features of the assessment space could be extended to the

instructional space (e.g., games, videos, tasks) and the setting in which the child is observed (e.g., the classroom) or any other object or element that is hypothesized to affect a child's learning (Baker, 2015a, 2015b; Baker et al., 2015; Chung & Parks, 2015; Redman & Kennedy, 2017). When statistical analyses are conducted on the relationship between the features and performance, the results may identify potential growth areas for students, identify content areas amenable to instruction, and provide a method for comparability and prediction of student performance (Baker, Cai et al., 2015; Baker, Madni et al., 2015).

For games, we use a standardized set of features that describe the interaction opportunities of a game (Chung & Parks, 2015; Redman & Kennedy, 2017). These are features related to the type of input the player is allowed to submit to the game, the kind of feedback provided to the player by the game, and how the game presents the targeted constructs. This feature set is based on media research, instructional practices, and CRESST's experience creating and studying educational games. Some feature set iteration may be necessary during analysis as salient features emerge that were not initially included in the list. Feature set iteration generally occurs at the beginning of the analysis process before the bulk of the games have been rated. However, revision of the feature list may be warranted even in later stages of the analysis if a salient novel feature is discovered in a new game or there is cause to amend a definition to more accurately and reliably rate the games. Whenever the feature list is revised, all already analyzed games must be re-rated with the new features and definitions in mind. At its core, the process endeavors to develop a stable and inclusive feature list that can be reliably applied across various games.

The utility of having a qualitative method of evaluating a game was examined by Redman et al. (2023). One objective of Redman et al. was to investigate whether games classified as having more learning potential, compared to games classified as having less learning potential, would show in-game performance gains (presumably due to learning of the content). An initial feature analysis of 15 existing PBS KIDS games with high data quality was conducted, a process that yielded 12 games that had alignment of learning goals, gameplay, and measurement potential. A more in-depth feature analysis was then done using the features in Appendix A. The analysis resulted in three games classified with a learning potential of not likely or less likely, and four games classified as likely. Data collection occurred over five months and analysis was conducted with data from five of the games.

Learning was modeled with a two-timepoint latent variable model where the inputs to the model were gameplay performance indicators. The two games classified as *not likely* or *less likely* to have a learning potential had a change in latent ability scores of .08 and .12, with both games having effect sizes of 0.07. The two games classified as *likely* had a change in latent ability score of 0.30 and 0.42 (the third game's model did not converge), with effect sizes of 0.59 and 0.56, respectively.

These results are consistent with the idea that the qualitative rating of games using learning-focused features in Appendix A can detect a game's learning potential a priori. Stated another way, the features in Appendix A—particularly the instruction and feedback features—provide guidance on the learner-system interactions that may be sensitive to learning.

Challenge 2: Identifying Potential Game-Based Indicators and Developing Algorithms to Derive Those Indicators From the Atomic Units of Evidence

The crux of high-quality learning process data is challenge 2. Challenge 2 arises because what constitutes data is wholly defined by the software that is implemented to capture and record the data. Early decisions about what behavior to log and at what granularity, when to log it, and what format to store the data can substantially impact downstream processes.

Deciding on What Behavior to Log

The first step is necessary but insufficient in extracting meaning from fine-grained learner-system interaction data. Defining what constitutes an atomic unit is crucial for subsequent analysis. Too fine-grained data logging (e.g., cursor movements) may result in unwieldy data and require extensive post-processing coding to reduce it to a usable form. Too coarse-grained data (e.g., logging only the solution submission) may omit highly informative behavior of learners' decisions and choices and preclude any possibility of examining fine-grained process questions. For example, in *Save Patch*, if learners' adding ropes to signposts were not recorded, or if the denominators of the rope and the current denominator in the signpost were not recorded, then it would be unlikely that any misconceptions could be identified through gameplay.

Another example is related to the fidelity of experience. For instance, game instructions are often presented through tutorials describing the game goals and game mechanics. If the tutorial is made skippable (e.g., by clicking a dismiss button)—as is often the case in games to maintain an enjoyable experience—then some players may skip the tutorial and later in the game not know what to do. The gameplay of these players may differ significantly from that of those who went through the tutorial. However, if the learner-system interactions on the tutorial were not logged, we would have no way of knowing whether the tutorial was skipped. Knowing whether learners skipped the tutorial allows us to describe learners more precisely, conduct more refined analyses about learning, and inform developers about usability issues.

Another important decision point when deciding what to log is the sampling policy. When and how frequently to sample the behavior can have essential post-processing implications. For example, logging that uses continuous sampling (e.g., 128 samples per second) may be appropriate for situations where the behavior is continuous, such as when measuring learners' fine-grained motor skills (e.g., see Nagashima et al., 2009). However, based on our experience attempting to make sense of data from learning games using continuous sampling of the entire game world, we think a more effective sampling scheme is event-based sampling that uses learners' overt behavior to trigger the logging of an interaction. Continuous sampling is simple to implement but records the state of the game world at fixed time intervals. This type of data requires substantial coding to extract events of interest. In contrast, event-based sampling requires modifying the game software and is more complex because decisions about what to log and at what granularity are necessary. Event-based sampling focuses on what interactions may be of interest a priori.

Finally, a related issue is the structure of the data logged. As a practical matter, the logging format can influence the amount of programming effort required to extract data. Design decisions about the data format include expressiveness, compactness, and with large datasets, computing and storage resources. Chung (2015) presents some guidelines on the design and implementation of telemetry, as do others (e.g., Hao et al., 2016).

Developing Indicators Rests on Algorithms and Coding

The rationale for capturing learner-system interactions is to use these interactions as inputs to algorithms to derive indicators of learning processes and outcomes. The indicators themselves can be used directly or as inputs to measurement models of higher level constructs (See Figure 4). The practical question is how to transform a sequence of low-level behaviors into indicators that reflect learners' thinking.

Chung et al. (2023) provide a concrete example of this challenge, highlighting why we assert that indicator development is, in fact, algorithm development and coding guided by a learning and measurement perspective. The game targeted computational thinking, and the measurement question was how to evaluate a player's toy design, which is composed of four parts, in a way that accounts for the outcome (whether it satisfies the design requirements or not) and reflects the problem-solving and debugging processes.

In the game, the player is tasked with designing a toy that meets certain specifications. The player is given various toy parts to use during the building phase, and then in the test phase the player can test the toy to see if it meets the requirements. If the toy does not meet the criteria, the player has to adjust the toy's design.

Our approach was to develop a method for comparing the four components of the learner's toy design to reference solutions. This method allows for the computation of several similarity scores for any toy design: a composite score for the overall design and scores for each toy component. We reasoned that the quality of the design is indicative of the learner's problem-solving and debugging process outcomes. Computing overall and component scores for each attempt allows for tracking progress over time.

Appendix B shows a snippet of an indicator design document we developed for the PBS KIDS game *Toy Maker*, detailing how to compute the overall composite score and component scores for each toy part. Appendix B shows that establishing a common vocabulary is the first step. Measurement considerations in light of the game design are a critical next step. The general requirements for the indicators are identified, such as being able to measure changes in players' responses (in our case, determining how close a player's design is to a solution given the game

presents a problem-solving task), being able to compare players in a consistent way (given that players may approach the game in different ways), and being able to differentiate players who use different problem-solving strategies as reflected in different but acceptable game designs.

To achieve the measurement requirements, we examined the solution space for the most critical elements (i.e., what constitutes a valid design, what contributes to a valid design, and how we can operationalize the detection of a valid design). Our solution was to establish a set of rules that reflect whether the player satisfied a specific condition for a particular part as well as the parsimony of the solution. The use of component rules allows for flexibility in terms of how the rules can be weighted for scoring purposes (or not) and the ability to describe players' performance for each toy (e.g., reporting which rules were met or not for each toy).

We included Appendix B to provide insight into the actual indicator development process used in a game designed for 6-year-old children. We wanted to emphasize that algorithm development and coding are essential parts of indicator development, which is unavoidable when working with interactive data in digital systems (coding was required to derive indicators for each game example presented in this chapter). The coding level of effort is influenced tremendously by which learner-system interactions are logged, the game's complexity, the extent to which interactions can be evaluated, and the availability of reference structures for comparison purposes.

Addressing Challenge 2

The most effective way to meet challenge 2 is to adopt a measurement perspective centered around learning (Baker, 1997) with a focal point on the relation between fine-grained behavioral interactions and attendant cognition. As an intellectual tool, a measurement perspective naturally leads to two fundamental questions: What is to be measured? How is it to be measured? Indicator development involves, in the end, an algorithm to be coded to operate on fine-grained behavioral data. Thus, reasoning about how a task design shapes a learner's behavioral responses can reveal the likely cognitive demands of a task. Likewise, reasoning about how a user-interface design enables or constrains learners' ability to express their thinking can reveal the evidentiary value of a learner-system interaction.

Addressing the "what-to-measure" and "how-to-measure" questions often results in an iterative process. For example, desiring to measure problem-solving leads to more questions and increasing definitions about the cognitive demands: Problem-solving about what? What types of problem-solving can be expected of learners (e.g., trial-and-error vs. means-ends)? Under what conditions are learners solving problems (e.g., closed-ended or open-ended problems, resource availability, type of feedback), and with what kinds of learners (e.g., degree of prior knowledge)?

A similar definitional process occurs during indicator development. Given the task design, what learning processes are learners likely to use during the task? How are learning processes expressed through the user-interface elements? For example, terms like "performance," "learning," and "proficiency" can be characterized in different ways and are unlikely to be immediately operationalized. Thus, deconstructing these terms into increasingly more precise definitions will help identify learner-system interactions that can satisfy the definition in the context of the task and cognitive demands. The definitional process may also reveal whether the interactions can be evaluated and used directly or need to be transformed, combined, or evaluated in the context in which the action occurred. This degree of detail is necessary because at some point, code will need to be written to transform the raw behavioral data into an indicator value. As was realized over 50 years ago in software engineering, software development projects are more likely to succeed when clear, precise, and complete requirements are documented (Brooks, 1975).

Challenge 3: Gathering Validity Evidence

The crux of credible indicators is the third challenge: validity evidence. Challenge 3 is important because it involves a potentially complex set of transformations performed on the learner-system interaction data to derive indicators of learning processes. The process of transitioning from the event layer to the indicator layer in Figure 4 is realized by algorithms and code, and thus, the transformations may not be easy to inspect and evaluate with respect to the relation between learners' behavior and presumed learning processes. Furthermore, the "degrees of freedom" in learner-system interactions are mediated by the design of the task, necessitating careful attention to learners' responses. Learner-system interactions are highly dependent on the design of the software and user interface—the universe of learners' behaviors is defined by the actions allowed by the user interface.

Addressing Challenge 3

AERA et al. (2014) define standards for validity and the various forms of validity evidence (pp. 23). Sireci and Benítez (2023) provide concrete examples of validation and validity evidence in the context of educational testing. When validation concepts are applied to game-based indicators, both qualitative and quantitative methods can be used for evidence gathering. In this section, we present examples drawn from our own work to illustrate the validation process. In general, our objective is to critically evaluate the extent to which the information encoded in the game-based indicators captures the relevant and meaningful aspects of the target constructs. Qualitative strategies include examination of the game design, game mechanics, and gameplay. Quantitative strategies include the examination of bivariate relations between various game-based indicators and external tests targeting the same construct and, most recently, joint validation of game-based indicators against other outcomes.

Qualitative Approaches

Because learner-system responses are highly dependent on the design of the software and user interface, the game design, game mechanics, and gameplay are all examined. For example, the game design and game mechanics undergo a feature analysis as described in the previous section, *Addressing Challenge 2: Feature Analysis*. Additionally, observation of learners' actual gameplay is critical to explaining unusual learner-system interactions as well as uncovering potential issues with the data. Learners are allowed to play as naturally as possible with essentially no help or intervention from the researchers. This type of observation provides information on where in the game players get stuck and what aspect of the game is confusing (e.g., not understanding the goal; incomplete understanding of the game controls and interface elements; unclear, ignored or missed directions, help, hints, and feedback).

Another qualitative approach is what we refer to as "reverse response process validation." By "reverse," we mean developing game-based indicators using extant microgenetic studies (e.g., Metz, 1993; Siegler, 2007) where we assume that the research base, findings, and theory impart validity. Microgenetic analysis densely samples observations of how learners use their knowledge (or not), how they develop and discover strategies (or not), and how learners transition toward mastery within a subject. These observations are of fine-grained, real-time behaviors that likely covary with the unfolding learning process. Microgenetic studies typically have a line of research that includes theoretical frameworks and prior findings.

For example, we developed a game-based indicator algorithm based on Metz's (1993) microgenetic analysis for the PBS KIDS' game *Pan Balance* (https://pbskids.org/sid/games/pan-balance). In her study, Metz examined how preschoolers built and refined their procedural and diagnostic knowledge of weight and the use of a pan balance. Metz identified patterns of misconceptions that accompanied changes in knowledge. One such misconception, called "higher is heavier," occurs when children mistakenly interpret the higher side of the pan as containing the heavier object. We found that children exhibiting this misconception tended to show less raw change from the pretest to the posttest (Chung & Feng, 2024; Redman et al., 2018).

Quantitative Relations to Other Measures

A conventional way to gather quantitative validity evidence is to evaluate the relationships between game-based indicators and externally validated measures, given that both the externally validated measure and the game share the same set of cognitive demands. One important function of the external measure is to serve as a reference measure of knowledge and skills. The use of experimental conditions and subgroups allows testing of game-based indicators to check whether the indicator values reflect expected directions. For example, an essential property of a measure when learning is involved is instructional sensitivity (Baker, 1997). Assuming instruction was effective, the indicator value prior to instruction should be lower compared to the indicator value post-instruction. Similarly, the indicator value should be higher for learners who already possess the target knowledge or skill compared to those who do not possess the target knowledge or skill. If these relations exist with the external measure, then a similar pattern should also exist with the game-based indicators. Such a pattern of results would be strong validity evidence. Table 4 summarizes the different kinds of comparisons that can be done to provide validity evidence.

Table 4.
Summary of Potential Validity Evidence

General analysis		Potential validity evidence
1.	Correlation between external pretest scores and GBIs (gameplay on early rounds).	GBIs are sensitive to preexisting skills and knowledge.
2.	Correlation between external posttest scores and GBIs (gameplay on later rounds).	GBIs are sensitive to learned (or existing) skills and knowledge.
3.	Correlation between the gain scores of the external measures (posttest–pretest) and gain scores of the GBIs (later rounds–early rounds).	GBIs are sensitive to the degree of learning of skills and knowledge.
4.	Subgroup analyses: Compare GBIs of players who *learned* to players who did not learn over the course of the intervention. "Learned" is defined as a positive pretest to posttest gain on the external measure.	If the GBIs of players who learned (vs. players who did not learn) show differences in the expected direction (e.g., show more use of productive processes, less use of nonproductive processes, less errors), then this result would suggest that players who learned use more productive processes than players who did not learn.
5.	Subgroup analysis: Compare (early round) GBIs of players who scored high on the *pretest* external assessment to GBIs of players who scored low.	If the GBIs of players who have high pre-existing skills and knowledge (vs. players who have low pre-existing skills and knowledge) show differences in the expected direction (e.g., show more use of productive processes, less use of nonproductive processes, less errors), then this result would suggest that players who have higher skills and knowledge use more productive processes than players with lower skills and knowledge.
6.	Subgroup analysis: Compare (late round) GBIs of players who scored high on the *posttest* assessments to GBIs of players who scored low.	If the GBIs of players who have high skills and knowledge at the end of the game (vs. players who have low skills and knowledge) show differences in the expected direction (e.g., show more use of productive processes, less use of nonproductive processes, less errors), then this result would suggest that players who have higher skills and knowledge use more productive processes than players with lower skills and knowledge.

Note. GBI = game-based indicators.

Chung and Feng (2024) present game-based indicator validity evidence for various games and additional examples exist involving different games and interactive systems, external measures, ages, interventions, and type of process data (e.g., Chung & Baker, 2003; Chung et al., 2002; Choi, Parks et al., 2021; Choi, Suh et al., 2021; Feng, 2019; Feng & Cai, 2024; Kerr, 2014; Kerr & Chung, 2012; Nagashima et al., 2009; Redman, Chung, Feng et al., 2020a; Redman, Chung, Griffin, & Parks, 2020b; Redman et al., 2018, 2021, 2023; Teng & Chung, 2025).

Joint Modeling of Game-Based Indicators and Validation of Multiple Sources of Evidence

Advances in methodology now allow the validation of multiple game-based indicators that collectively describe a single phenomenon of interest. Information encoded in these indicators can be integrated into a larger, reliable system for measuring performance and learning.

Correlational analysis and multiple linear regression, two of the most used analysis techniques (Zhu et al., 2023), fall short when the goal is to analyze multiple indicators targeting the same construct simultaneously, jointly model these indicators with other measures, or examine relationships without aggregating variables to the player level.

Paradigms of latent variable modeling, such as item response or factor models, offer flexible means to accommodate a wide range of analytic decisions (Skrondal & Rabe-Hesketh, 2004). Item response theory, multilevel modeling, and diagnostic classification modeling have been applied to analyzing one or more data sources collected in game-, simulation-, or computerized task-based research (e.g., Choi, Suh et al., 2021; Feng & Cai, 2024; Liu et al., 2018; Reese et al., 2015).

By examining the relationship between a game-based latent factor, measured by a set of game-based indicators, and one or more assessment-based latent factors, measured by sets of items, we can gauge the extent to which learners actions in an interactive system, such as a learning game, correlate with learning outcomes. From a validation perspective, this approach is tantamount to being able to validate multiple game-based indicators against external assessment item responses. Feng and Cai (2024) demonstrated the analytic benefits of jointly modeling diagnostic indicators, derived from gameplay process data for each

in-game task, with traditional pretest-posttest item response data collected in game-based evaluation research.

A benefit for learning research is the ability to connect students' interactions—such as patterns of misconceptions—in a low-stakes, game-based setting, with changes or hindered changes in their educational outcomes that are typically valued in higher stakes settings. One implication of being able to validate diagnostic indicators, whether through a model-based approach or others, is the ability to use these indicators to monitor learner-system interactions and provide feedback that is both relevant and timely.

The qualitative and quantitative approaches described yield a range of validity evidence, from response process evidence to statistical evidence (AERA et al., 2014), and support *Principle 3* (assessment design). Collectively, these approaches are intended to identify the underlying reasons driving learners responses and to test for patterns of relations consistent with expectations (Sireci & Benítez, 2023). The evidence collected is useful for adjusting the design of the task, particularly when learners respond to the game in unexpected ways or when the game-based indicators reveal unexpected relationships. Such incongruence, left unaddressed in the game design or unidentified, becomes pernicious at the analysis and interpretation stage. Behavior that may appear productive may actually be due to reasons entirely unrelated to the target knowledge or skill, leading to biased results and improper inferences.

Discussion

One reason for using technology-based tasks for measurement purposes is that with judicious task design, software can be developed to elicit from a learner complex cognitive processes (e.g., problem-solving, reasoning, creativity, self-regulation, adaptivity, metacognition, collaboration) in the context of some content domain (Baker, 1997; Baker et al., 2016) and offer a more scalable option than other modes such as hands-on performance tasks. A second reason is that the task can be instrumented to automatically track both the process a learner uses to complete a task and the performance outcome of the task. These two capabilities enable the development of rich and highly interactive tasks, scalable administration, unobtrusive behavioral observations, and automated scoring of learner processes and task outcomes.

Although the first reason is widely accepted, as evidenced by the inclusion of technology-enabled tasks in large-scale testing programs (e.g., NAEP and PISA), the second reason—the promise of process data—continues to face challenges (Feng & Cai, 2024; Lindner & Greiff, 2023). To fully realize assessment in the service of learning, not only to understand what learners know and can do, but also to measure the learning processes students are using (or not using) and using that information for instructional purposes, the shortcomings surfaced by Lindner and Greiff (2023) and others (e.g., Chung & Feng, 2024) need to be addressed. Advances are required to move the indicator development process from an artisan activity to an engineering process. The opaqueness of the indicator development and the direct impact algorithms and coding have on the indicator's value led Chung and Feng (2024) to assert that advances are needed in three areas: traceability, interpretability, and algorithm generalizability. *Traceability* refers to the ability to trace how the raw interactions are processed and transformed (e.g., filtered, aggregated, recombined) into a quantitative value. Given an algorithm, *interpretability* refers to how one interprets the value produced by an algorithm—the meaning ascribed to the indicator in light of the assumptions, constraints, and transformations encoded in the algorithm. *Algorithm generalizability* refers to how well an algorithm, based on a theory, encodes the rules and conditions described or predicted by the theory. Algorithm generalization occurs by applying the algorithm (with modification to adjust for task-specific surface features) to generate indicators on tasks that may differ in format, mechanics, content, or even learning goals.

The idea of collecting interaction data in digital systems is not new. What is new is that we have conceptualized learner-system interactions as an observation to explicitly support measurement and thus a concomitant focus on validity. Interestingly, a side effect of the challenges in indicator development may be an increased focus on the meaning of learner-system interactions. To develop software, detailed specifications of what to produce is needed. This demand for detail and definitions may increase awareness of how learner-system interactions represent evidence of the target knowledge or learning process.

In this chapter, we have attempted to illustrate how well-designed instructional opportunities in interactive systems provide measurement opportunities. These opportunities can result in what we call measurement without testing: Learner-system interactions that are designed to support students' learning are, by

definition observable, and we believe they carry the most relevant information about students' learning. Observing learners' interactions in digital systems, whether games or simulations, is still the only scalable method for observing large numbers of students compared to other forms of observation, such as video recording, audio recording, eye tracking, EEG, fMRI, and physiological and motor monitoring. If we can observe what learners are doing as they do it and accurately determine why, then that capability may help move us toward tailored, adaptive, and individualized learning for all students.

References

Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(5), 883–901.

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

Anderson, J. R. (1982). Acquisition of cognitive skills. *Psychological Review, 89*(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge University Press.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247–254. https://doi.org/10.1080/00405849709543775

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

Baker, E. L. (2015a, April 16–20). Feature analysis as a technology design and evaluation tool. In G. K. W. K. Chung (Chair), *Design issues regarding the use of games and simulations for learning and assessment* [Roundtable]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

Baker, E. L. (2015b, April 16–20). *The design and validity of new assessments: Windows on architecture, art, & archaeology* [Invited speaker session]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

Baker, E. L., Cai, L., Choi, K., & Madni, A. (2015, June 22–25). *Functional validity: Extending the utility of state assessments* [Conference session]. 2015 National Conference on Student Assessment, San Diego, CA, United States.

Baker, E. L., Chung, G. K. W. K., & Cai, L. (2016). Assessment gaze, refraction, and blur: The course of achievement testing in the past 100 years. *Review of Research in Education, 40*, 94–142.

Baker, E. L., Madni, A., Michiuye, J. K., Choi, K., & Cai, L. (2015). *Smarter Balanced Assessment Consortium: Mathematical reasoning project quantitative analyses results: Grades 4, 8, and 11*. Smarter Balanced Assessment Consortium.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007–466). U.S. Department of Education, National Center for Education Statistics. https://eric.ed.gov/?id=ED497845

Brooks, F. (1975). *The mythical man-month: Essays on software engineering*. Addison-Wesley Publishing Company.

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

Center for Advanced Technology in Schools, & CRESST (2013a). *CATS knowledge and item specifications: Rational number equivalence* (Revision 10/25/13). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools, & CRESST (2013b). Save Patch *tutorial and game level design: RN1.6*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

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.

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., 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*(6), 669–684.

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* (pp. 327–354). Springer International Publishing. https://doi.org/10.1007/978–3-031–56365–2_17

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, *15*(3–4), 463–494. https://doi.org/10.1016/S0747-5632(99)00032-1

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

Chung, G. K. W. K., Redman, E. J. K. H., & Choi, K. (2023). *Wombats analytics evaluation–Final plan* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

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]. American Educational Research Association Annual Meeting, 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]. American Educational Research Association Annual Meeting, Virtual Conference, United States.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, *26*(5), 1024–1033. https://doi.org/10.1016/j.chb.2010.03.003

Feng, T. (2019, April 5–9). *Using game-based measures to assess children's scientific thinking about force* [Poster presentation]. American Educational Research Association Annual Meeting, 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*, 12396. https://doi.org/10.1111/jedm.12396

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

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills*. OECD. https://doi.org/10.1787/e5f3e341-en

Gordon, E. W. (1970). Toward a qualitative approach to assessment. *Report of the Commission on Tests, II. Briefs* (pp. 42–46). College Entrance Examination Board.

Gottman, J. M., & Notarius, C. I. (2000). Decade review: Observing marital interaction. *Journal of Marriage and Family*, *62*(4), 927–947. https://doi.org/10.1111/j.1741-3737.2000.00927.x

Greer, R. D., & McDonough, S. H. (1999). Is the learn unit a fundamental measure of pedagogy? *The Behavior Analyst*, *22*(1), 5–16. https://doi.org/10.1007/BF03391973

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). Taming log files from game/simulation-based assessments: Data models and data analysis tools. *ETS Research Report Series*, *2016*(1), 1–17. https://doi.org/10.1002/ets2.12096

Heitz, R. P. (2014). The speed-accuracy trade-off: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnins.2014.00150

Janlert, L.-E., & Stolterman, E. (2017). *Things that keep us busy: The elements of interaction*. MIT Press. https://doi.org/10.7551/mitpress/11082.001.0001

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

Kennedy, G. E. (2004). Promoting cognition in multimedia interactivity research. *Journal of Interactive Learning Research*, *15*(1), 43–61. https://www.proquest.com/docview/1468384849/citation/131F6A71935242F4PQ/1

Kerr, D. S. (2014). *Into the black box: Using data mining of in-game actions to draw inferences from educational technology about students' math knowledge* [Unpublished dissertation, ProQuest No. 3613716]. University of California, Los Angeles.

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1), 144–182.

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

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology, 9*(1372).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Metz, K. E. (1993). Preschoolers' developing knowledge of the pan balance: From new representation to transformed problem solving. *Cognition and Instruction, 11*(1), 31–93. https://doi.org/10.1207/s1532690xci1101_2

Mislevy, R. J., Riconscente, M., & Corrigan, S. (2015). Evidence-centered assessment design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 40–63). Routledge. https://doi.org/10.4324/9780203102961

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.

National Center for Education Statistics. (2012). *The nation's report card: Science in action: Hands-on and interactive computer tasks from the 2009 science assessment* (Report No. NCES 2012–468). Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2009/2012468.pdf

National Center for Education Statistics. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study* [White Paper]. Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. U.S. Department of Education.

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Erlbaum.

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

Organisation for Economic Co-operation and Development (OECD). (2023). *PISA 2025 Learning in the digital world framework (second draft)*. OECD Publishing. https://www.oecd.org/media/oecdorg/satellitesites/pisa/PISA%202025%20Learning%20in%20the%20Digital%20World%20Assessment%20Framework%20-%20Second%20Draft.pdf

Ostrov, J. M., & Hart, E. J. (2013). Observational methods. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 285–303). Oxford University Press.

Plass, J. L., Schwartz, R. N., & Heidig, S. (2012). Interactivity in multimedia learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1615–1617). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_1848

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. (2020a). *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.

Redman, E. J. K. H., Chung, G. K. W. K., Griffin, N., & Parks, C. B. (2020b). *Social-emotional learning games analytics validation study design* (Final deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

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.

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 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. https://doi.org/10.4324/9780429282201

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

Redman, E. J. K. H., & Kennedy, A. A. U. (2017). *Feature analysis framework for Measure Up* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics: Actionable measures of multi-dimensional learning trajectories. *British Journal of Educational Technology, 46*(1), 98–122.

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*(2), 257–266.

Siegler, R. S. (2007). Microgenetic analyses of learning. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (6th ed., pp. 464–510). Wiley.

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, *35*(3), 217–226. https://doi.org/10.7334/psicothema2022.477

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901–926.

Teng, K., & Chung, G. K. W. K. (2025). Measuring children's computational thinking and problem-solving in a block-based programming game. *Education Sciences, 15*(1), 51. https://doi.org/10.3390/educsci15010051

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

Vendlinski, T. P., Delacruz, G. C., Buschang, R. E., Chung, G. K. W. K., & Baker, E. L. (2010). *Developing high-quality assessments that align with instructional video games* (CRESST Report 774). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles. http://files.eric.ed.gov/fulltext/ED512655.pdf.

Webb, N. M. (1983). Predicting learning from student interaction: Defining the interaction variables. *Educational Psychologist*, *18*(1), 33–41. https://doi.org/10.1080/00461528309529259

Williams, M. D., & Dodge, B. J. (1993). *Tracking and analyzing learner-computer interaction*. Proceedings of Selected Research and Development Presentations at the Convention of the Association for Communications and Technology. https://eric.ed.gov/?id=ED362212

Young, M. F., Kulikowich, J. M., & Barab, S. A. (1997). The unit of analysis for situated assessment. *Instructional Science*, *25*(2), 133–150. https://doi.org/10.1023/A:1002971532689

Zhu, S., Guo, Q., & Yang, H. H. (2023). Beyond the traditional: A systematic review of digital game-based assessment for students' knowledge, skills, and affections. *Sustainability*, *15*(5), Article 4693. https://doi.org/10.3390/su15054693

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/a000748

Appendix A:

Learning-Related Features and Definitions

Category Feature	Feature Description
Learning-Goal Alignment	
Learning Goal Aligned with Gameplay	Game requires players to access and use the targeted learning goal(s) in order to play the game. Gameplay is not tangential to the targeted learning goals and the player must understand the targeted learning goals in order to be successful in the game. When the gameplay and learning goals are *not* aligned, a player is able to succeed at the game without demonstrating understanding of the learning goals.
Learning-Goal Explanation	
Explanation Provided	Game provides introduction to or background information for the targeted learning goal(s) before or during gameplay (this information is distinct from feedback). This information *may* be presented in an in-game tutorial, but the presence of a tutorial does not automatically imply an explanation of the target learning goal(s).
Type of Progression	
Fixed Learning-Goal Complexity	Gameplay tasks/rounds/levels do not change in complexity over the course of the game. This does not mean that the tasks are exactly the same over the course of the game, just that they do not change in level of complexity or difficulty.
Increasing Learning-Goal Complexity	Gameplay tasks/rounds/levels presented become more complex or "hard" as the player advances in the game. Learning-goal complexity refers to the targeted learning goals that are present in the game. Learning-goal complexity may increase by way of the inclusion of additional learning goals as the game progresses.
Adaptive Game Progression	The game serves up tasks/rounds/levels based upon player performance. This means that the level order is dependent on player input and will not necessarily be the same for each player.
Fixed Game Mechanic Complexity	Gameplay tasks/rounds/levels do not change in game mechanic or user interface complexity over the course of the game. This does not necessarily mean that tasks are exactly the same in terms of game mechanics over the course of the game, just that they do not change in level of complexity or difficulty. This is entirely independent of learning-goal complexity, which relates to the targeted learning goals.

Category Feature	Feature Description
Gameplay Type	
Judgment/Decision Making	Players are asked to make judgments/decisions based on their understanding of the target learning goal(s).
Input Submission	
Intentional Submission	Players must intentionally submit their answer/response to stimuli in the game. This may take the form of a submit button (like *Pan Balance* or *Meatball Launcher*). The point is that submission must be intentional as some games may allow players to manipulate the game space and automatically accept a correct response (whether the player means to submit it or not). A gamified assessment (in which a player must select a correct response from several items, similar to a multiple choice question) does not count as an intentional submission unless there is a further step to confirm the selected response is the intended answer.
Creative Submission	Game requires players to create something or perform an activity. This is different than a game that requires the player to select a correct object or response.
Instruction and Feedback	
Demo	The game provides a demonstration that explains and shows the gameplay to players by walking them through a task/gameplay. The demo may include directions about the target learning goals and/or gameplay. Most games begin with some sort of background information or gameplay directions; these do not count as a demo unless the player sees a demonstration of the gameplay. The demo cannot be interactive.
Tutorial Level	The game provides a tutorial that requires players to participate in a demonstration of how to play a level/task or how to manipulate specific elements of the game. The tutorial must be interactive (i.e., require player participation or input).
Demo Skip Option	Game allows players to skip the demo, if desired. Presence of a skip button does not necessarily indicate the presence of a demo. Some games allow players to skip the instructions or background story/information. Note: Demo skip option may only be present if a demo exists and is able to be skipped.
Individual Learning-Goal Presentation	If there is more than one learning goal targeted by the game, it presents the learning goals individually (not at the same time).
Modal Feedback	Modal feedback requires players to attend to the feedback while it is being given. Gameplay and game interactions are disabled while feedback is being delivered. This means that players are unable to skip feedback.

Category Feature	Feature Description
Audio-Visual Feedback	Feedback is provided both in audio and visually (text or other visual clue).
Correct Answer Acknowledgment	Feedback or acknowledgment of a correct input is provided without elaboration about why it is correct (e.g., "good job!" or "that's right" or another audio or visual clue that indicates success).
Correct Answer Elaboration	Feedback acknowledges the correct input but ALSO explains why it is correct by elaborating on the target learning goal (e.g., "you are right, that block is taller than the other ones"). Elaboration does not need to occur for each round of feedback, but should be marked if it is present at any point in the game.
Incorrect Answer Acknowledgment	Feedback or acknowledgment of an incorrect input is provided without elaboration about why it is incorrect (e.g., "try again" or "that's not right" or another audio or visual clue that indicates an incorrect input).
Incorrect Answer Elaboration	Feedback acknowledges the incorrect input but ALSO explains why it is incorrect by elaborating on the target learning goal (e.g., "that's not right, that block isn't taller than the other ones"). Elaboration on the game mechanic (e.g., "those dinosaurs are not in the right order") does not count as Incorrect Answer Elaboration. Repetition of the task prompt after signaling an incorrect answer does NOT count as elaboration if it does not also include some explanation of what was incorrect. Elaboration does not need to occur for each round of feedback, but should be marked if it is present at any point in the game.
Graduated Feedback	Feedback becomes progressively more explicit or helpful as more errors are made by the player. This includes removal of incorrect answer options for selected response tasks, hints about the correct answer or how to complete the task, and other means for helping the player successfully advance in the game.

Category Feature	Feature Description
Constructive Processes	
Prediction	Game asks the player to predict outcome(s) based upon given information or game states. Usually prompts will ask, "what will happen next?" "What will happen if…" or ask the player to manipulate variables in the game space to effect a certain outcome. This does not apply to games that just ask a player to select a correct object or response (in the vein of a selected response assessment).
Reflection	Game asks the player to explicitly reflect on their answer or input (e.g., compare it to prediction/hypothesis, think about whether something worked, etc.).
Questioning	Game asks rhetorical questions about the target learning goal(s). This occurs (more often) in exploration games where the questions are rhetorical because the player is not required to answer them as part of gameplay.
Debugging/Correction	Game asks the player to correct or refine input based upon feedback. For example, if the player builds something that is unsuccessful, and the game asks the player to improve it so that it works better, this is debugging/correction. However, the game must present the player with their original creation/submission to fix, and not have them start again/redo their creation/submission.

Appendix B:

Indicator Design Document

Note: This appendix is an excerpt from an indicator design document for a game in a current study. Identifying names of the game has been renamed to generic labels.

Definitions

The following terms are used throughout this documentation. They are used to establish a shared language when we discuss various game-based indicators and the algorithms used to implement the indicators.

- **Toy type:** A term used to refer to one of the three types of toys that players can make in the game. These types are: Toy Type 1, Toy Type 2, and Toy Type 3.
- **Design category (category):** A term used to refer to the part, the color, the sizing, or the power/battery of a toy design.
- **Task (in-game task, game task):** A term used to replace the typically used "game level" to avoid possible confusion with downstream statistical analyses, where the term "level" means something distinctly different from a game level (e.g., in multilevel modeling, a level refers to an aggregation level—item level, student level, classroom level, school level). In the game, a task is the player making a toy. There are a total of nine tasks in the game, three tasks per type of toys—Toy Type 1, Toy Type 2, and Toy Type 3—that players can make.
- **Level:** Not used in this document. In this document, "game level" is referred to as "task" or "in-game task" or "game task."
- **Rule**: The set of conditions that satisfy part of all of the criteria for a given toy, task, and toy component (part, size, color, or power). A task may have multiple rules that if all are satisfied, indicate the player has a solution for the task.
- **Attempt:** The window of gameplay that starts with selecting (or reselecting) parts and modifiers of a toy, confirming the toy design, building the toy, testing the toy, observing the results of the testing, and ends with trying again if the testing fails. Each task can have more than one attempt.
- **Confirming design:** This refers to when a player clicks the right arrow button to confirm their toy design (before building).
- **Testing design:** This refers to when, after the toy is built, a player clicks the test button to test their toy design and observe if the design passes the test by meeting all criteria for a given task.

- **Player's confirmed toy design (player's design):** This refers to a player's confirmed toy design that is then used to build the toy.
- **Task solution:** This refers to a pre-specified compact solution for a task in the game.

Compact Solution Per In-Game Task

Rationale and Context

The goal of having a set of compact solutions is to facilitate analytics. By "compact," we mean the most parsimonious (also see the second section named "properties of a compact solution"). By "facilitate analytics," we mean the following activities, most of which are concerned with the development of indicators that describe and differentiate players' in-game performance or progress:

1. The development of indicators that gauge some kind of changes in player response' quality or closeness to the goal state requires that we know one or more references that could represent the goal state. In other words, convergence [to] or divergence [from], or being productive or unproductive, is always gauged with respect to at least one reference.

2. We need to establish a consistent way for comparing performances between players. A compact solution, specified per task, is one such reference that enables between-player comparisons.

3. We might also be interested in differentiating players who complete the same task but with different strategies, assuming such differences would relate to players' varying degrees of problem-solving or debugging abilities. For example, for the same task, Player A could complete the task with the most compact solution, whereas Player B completes the task with redundant modifiers used.

How exactly we want to score players' performances, such as to what extent we are concerned about a player's design being fully correct or being the most parsimonious, can be decided when we develop the scoring algorithm.

Properties of a Compact Solution

A compact solution has the following properties that are applied to each of the three parts of a toy (e.g., a toy type 3 has three parts: a body, a door, and a decoration):

- There are no "unnecessary but not incorrect" modifiers added. If a part requires no size, color, and/or power modifier, leave the corresponding modifier section blank (e.g., an empty list).
- If no specific part is needed to pass the test, leave the part section blank (e.g., an empty list).
- For example, for one of the solutions of Toy Type 3 Task 1, a player can use any of the house bodies with one orange modifier and two small modifiers, can use any of the doors, and can use any of the decorations. Then for this solution and for each of the three parts, the part name is left blank (e.g., an empty list).
- If there is a specific part needed to pass the test, use the name of the specific part (e.g., a list with only one element, where the element is the part name).
- For example, for one of the solutions of Toy Type 3 Task 1, a player can use the first house body with one orange modifier and one small modifier, provided that the player also uses the fourth decoration, along with any of the doors. Then for this solution and for the body part and the decoration part, we specify the name of the first house body (*fairy*) and the name of the fourth decoration (*flag*).

Generally, we assume that for each part (p = 1, 2, or 3) of a compact task solution, we have specified the following information:

1. Names of the specific parts needed; leave it blank if it does not matter which specific part can be used.
2. Size modifier(s) needed; leave blank if there is no requirement about Part p's size.
3. The color modifier needed; leave blank if there is no requirement about Part p's color.
4. Whether the power modifier needs to be included or explicitly excluded; leave blank if there is no requirement about Part p's power inclusion or exclusion.

Set of Compact Solutions for Each Task

The following section provides details on how various indicators are derived from players' submitted responses, as well as the finalized design score, which will be used as the primary outcome for modeling. The final score incorporates multiple facets of performance, and it is the most sensitive to incremental changes in the levels that the game requires players to beat.

Algorithm: Converging to and Diverging From a Solution

1.1 Sub-construct

Begin to notice where errors exist in algorithms (sequences) and attempt to fix them (debugging) (e.g., a child recognizes that they need to put larger blocks at the bottom of a block tower to keep it from falling).

1.2 Overview

The basic approach to indicator developed for the game is to detect which rules are satisfied. For each toy, rules are formed for the combination of three parts, the color modifier, the size modifier, and the power modifier. Then a player's submitted solution is checked against the solution set, and each rule is evaluated to return true or false. The rules have hierarchy (e.g., Rule 0, Rule 1, Rule 2, or higher) such that the more the higher-level rules are satisfied, the closer the player is to a solution. The use of rules satisfied and unsatisfied (or met and unmet) also provides flexibility in terms of how the rules can be used for scoring purposes. An example is presented at the end to measure converging to or diverging from a solution set.

In the general approach, rules are defined to help determine whether a player's submitted solution attempt is getting closer to meeting the beat-round criteria. The different types of rules are:

- **Rule 0** is used to check if a player has added any unnecessary modifiers.
- **For Rule 1 and above,** the more rules that are satisfied, the closer the performance is to a solution and thus the closer the player is to beating an in-game task (i.e., making a toy that would pass the test). For example, each of the toy types (Toy Type 1, Toy Type 2, and Toy Type 3) in the game has three in-game tasks (Task 1 to Task 3).

1.3 Data Structures

- Set of possible compact solutions for each task of each toy type.
- A player's confirmed toy design (player's design).

1.4 General Approach

- Create a solution set of all possible solutions for each task of all toys (9 total)
  - 3 toy types
  - 3 tasks per toy type
- Given a player's confirmed toy design and a pre-specified task solution
  - Check if the player's design contains the correct part(s)
    - For example, a task in the toy type 3 section has three parts: the house body, the door, and the house decoration
  - Check if the player's design contains the correct color modifier for the right part, and if the correct color modifier is in the last position
  - Check if the player's design contains the correct size modifier(s)
    - Check for the any size modifier
    - Check if the size modifier is added to the right part
    - Check if the overall sizing effect (after executing all size modifiers) is in the correct direction
  - Check if the player's design contains a power modifier if required, or does not contain a power modifier when not required

1.5.1 Part Checking

Because for some tasks in the game there are certain combinations of toy parts that affect task completion, all parts of the player's toy design are evaluated jointly.

Binary Representation for Handling Part Interaction

To account for the interactions between parts, we use a binary representation to record the combination of all three parts in a player's toy design. Each part (Part01, Part02, or Part03) consists of five unique components, resulting in a total of 15 distinct components across the three parts. We represent the presence and absence of these components using a sequence of 15 binary digits (ones and zeros), with each digit corresponding to a specific component, arranged from Part01 to Part03, top to bottom. For each solution, a digit at index i is set

to 1 if the solution includes the component at that index, and 0 if the component is not required. We then check if the binary representation of the parts used in a players' response matches any of the binary representations associated with a compact solution. Note that for solutions involving OR relationships, multiple binary representations can be associated with the same solution.

If there is a part requirement:

- Rule 1. Check if the task-required part is the same as the player's selected part; return true if the rule is satisfied and false if not satisfied

If this is no part requirement (when the player can use any part):

- Rule 0. Return true.

1.5.2 Color Checking

If there is a color mod requirement:

- Rule 1. Check if the player added any color modifier when there is a color mod requirement, return true if the rule is satisfied and false if not satisfied.
- Rule 2. Check if the player added any color modifier to the right part, given Rule 1 is true, return true if the rule is satisfied and false if not satisfied.
- Rule 3. Check if the player added the right color modifier, given Rule 1 and Rule 2 are true, return true if the rule is satisfied and false if not satisfied.
- Rule 4. Check if the player added the right color modifier as the last color modifier, given Rules 1–3 are true; return true if the rule is satisfied and false if not satisfied.

If this is no color mod requirement:

- Rule 0. Check if a player added any color modifier; return true if the rule is satisfied and false if not satisfied.

1.5.3 Size Checking

If there is a size mod requirement:

- Rule 1. Check if the player added any size modifier when there is a size mod requirement; return true if the rule is satisfied and false if not satisfied.
- Rule 2. Check if the player added any size modifier to the right part, given Rule 1 is true; return true if the rule is satisfied and false if not satisfied.
- Rule 3. Check if the player overall achieved the same sizing direction (e.g., big or bigger), given Rule 1 and Rule 2 are true; return true if the rule is satisfied and false if not satisfied. This can be achieved by one of the following:

  a. Adding the right number of the right type of size modifier, or

  b. Adding modifiers that have the same sizing effect as the solution, but the number of modifiers added differs from the solution, or

  c. Adding modifiers such that their sizing effects balance out to be the desired sizing effect (e.g., if a task wanted one small modifier, and the player added two small modifiers and one big modifier, then overall the toy was down-sized once).

If this is no size mod requirement:

- Rule 0. Check if a player added any size modifier; return true if the rule is satisfied and false if not satisfied.

1.5.4 Power Checking

If there is a power mod requirement:

- Rule 1. Check if the player added the power modifier to the right part when the task asks for it, or if the player did not add the power modifier when the task explicitly did not ask for it; return true if the rule is satisfied and false if not satisfied.

If this is no power mod requirement:

- Rule 0. Check if a player added the power modifier; return true if the rule is satisfied and false if not satisfied.

1.6 Player Solution Evaluation

```
Given Task x of Toy Type y
  For each pre-specified solution for Task x
    For each player's confirmed toy design z (Attempt z):
    – Compute the number of rules met for the toy's part, size, color, and power
    – Compute the number of rules unmet for the toy's part, size, color, and power
```

The use of rules allows for flexibility in terms of how we weigh the rules for scoring purposes (or not), being able to describe players' performance by toy (e.g., reporting which rules were met or not for each toy).

The example below shows what is possible once we know which rules are met or unmet. One possible scoring rubric given the list of rules met or not met is presented in Section 2.6.1.

1.6.1 Relationship Between Indicators and Performance

Because Rule 1 through Rule 4 are defined in order of increasing difficulty of being met, meeting both Rule 1 and Rule 2 (or higher) indicates better performance than just meeting Rule 1. For instance, merely adding a color modifier without noticing which part needs the color or that the added color modifier's effect will be overridden by another color modifier added after it would only satisfy Rule 1, not Rule 2 and above.

1.6.2 Scoring Rubric

```
Given Task x of Toy Type y
For each pre-specified Solution s for Task x
  For each player-confirmed toy design:
   – If the task has a category requirement (part, color, size, or power),
     satisfying one rule within that category adds 1 point for that category.
   – For any category that is not required by the task, the rules do not count
     towards scoring. For example, Solution S may not require the player to add a
     color modifier, then all rules specified under Section 1.5.2 do not apply.
   – We assume that Rule 0 does not contribute to the scoring process.
   – For each category (part, color, size, or power), a category-specific score is
     computed.
   – An overall design score is the sum of four category-specific design scores,
     divided by the maximum number of points that can be earned for Solution S.
     – Overall design score = part score + color score + size score + power score
```

The rationale for not counting Rule 0 and the like is as follows. As long as the player does not add any part or modifier that leads them away from a solution, we do not penalize them for adding unnecessary modifiers or parts.

We can use the resulting score and changes in scores to gauge, over multiple attempts by one player, the extent to which the submitted toy designs converge to or diverge from a pre-specified solution, and to identify which specific rules are met or unmet.

Reflections on Reconceptualizing Assessment to Improve Learning

Stephen G. Sireci and Eric M. Tucker

The chapters in this Volume II of the *Handbook of Assessment in the Service of Learning* illustrate why the volume is subtitled *Reconceptualizing Assessment to Improve Learning*. This reconceptualization involves multiple facets of assessment from development through results reporting and highlights the shift from tests that measure status, to assessments that focus on engaging and supporting learners. The authors in this volume interrogate traditional models of developing tests, assessing students, and reporting information, and build upon that foundation to envision new approaches that expand assessment's capacity to inform and improve learning.

Rather than privileging one goal over another, this reconceptualization invites a broader framing—one where assessment serves both the advancement of learning and the need for fairness, evidence, and technical rigor. For if the purpose of assessment is to serve learners, test design, development, administration, and feedback must focus on that purpose. As these chapters illustrate, for assessments to truly serve learners, they must be flexible and multifarious, acknowledge the wide range of learners to be served, and embrace that diversity through design. Assessment design that serves learners will employ many methods such as game-based designs, portfolios, and personalization. The design will also include proven methods that draw from self-regulation principles, culturally responsive assessment, and learner engagement.

Validity in this paradigm is more than statistical tests; it encompasses evidentiary usefulness for teaching and learning, fairness, and the consequences of use—asking whether assessments *help* learners learn, and requiring evidence of that learning. For educators and test developers, that means engaging with learning

communities to design authentic tasks and report results in clear, diagnostic, and actionable ways—moving from opaque scales to feedback that informs next steps in a useful and usable manner. For policymakers, it calls for systems that privilege classroom-embedded assessment cultures and participatory co-design, and embrace policies that privilege understanding, support learning environments responsive to learner variation. The work ahead is clear: our field must invest in building tools, capacity, and enabling environments so assessments in the service of learning have the potential to be realized for learners, educators, and families.

The Promise of a More Humane, Learner-Centered Assessment

In the 1950s, while Professor Edmund W. Gordon served as an educational psychologist at the Pediatric Clinic of the Jewish Hospital of Brooklyn, he worked closely with Else Haeussermann, a special educator whose practice reshaped his understanding of assessment. Haeussermann was uninterested in sorting children by scores; she sought to understand *how* they learned and the conditions under which they succeeded (Gordon, 2020; Gordon, 2025). Together Haeussermann and Gordon studied learners' adaptive strategies—the moves children made when tasks were clarified, chunked, modeled, or connected to their experiences. Haeussermann's approach defied the conventions of test standardization and was deemed too labor-intensive, yet it represented a foundational model of assessment in the service of learning (Gordon, 2025). Their reports documented these patterns and their instructional implications, and Haeussermann translated the findings into concrete, individualized lesson plans (Gordon, 2020). That collaboration affirmed a principle Gordon never abandoned: in pedagogy, the primary purpose of assessment is to inform and improve learning, not merely to certify status (Gordon & Rajagopalan, 2016).

This journey is, at its heart, a commitment to honoring the whole learner (Armour-Thomas et al., 2019). It requires educators to become designers of rich learning environments, test developers to prioritize instructional and learning value in addition to psychometric elegance, and policymakers to foster systems that trust and invest in the professional expertise and capacity of educators.

Ultimately, the powerful and hopeful message of this *Handbook* is that the tools and frameworks we design are secondary to the humanistic vision that guides them. The final measure of any assessment's worth is not found in a score report, but in the confidence, curiosity, and competence assessment processes inspire in a learner.

Conclusion: Toward Assessment in the Service of Learning

Taken together, the insights from Volume II point to an education system where assessment serves learning. By acting on these principles, assessment will become a powerful engine for learning (Hattie, 2009). Volume III explores working examples and actionable blueprints for assessment in the service of learning.

References

Armour-Thomas, E., McCallister, C., Boykin, A. W., & Gordon, E. W. (Eds.). (2019). *Human variance and assessment for learning.* Third World Press Foundation.

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., & Rajagopalan, K. (2016). New approaches to assessment that move in the right direction. In E. W. Gordon & K. Rajagopalan (Eds.), *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.

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 with an interest 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 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. (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. W., 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. W., & Bridglall, B. L. (Eds.). (2006). *Affirmative development: Cultivating academic ability* (Critical issues in contemporary American education series). Rowman & Littlefield.

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*. 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 the 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. D., & 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 B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–135). 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 Eignor 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 Eignor 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 system-wide 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).

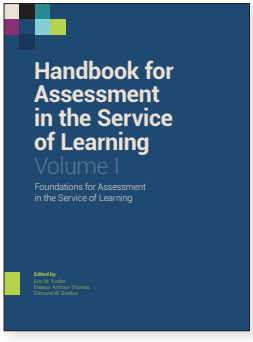
Alina A. von Davier is a researcher, innovator, and an executive leader with over 20 years of experience in EdTech and in the assessment industries. She is the Chief of Assessment at Duolingo, leading the Duolingo English Test research and development area. She is the Founder and CEO of EdAstra Tech. 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).

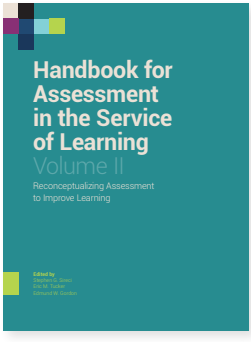
Kevin Yancey is a Senior Staff AI Researcher at Duolingo, leading the engineering and AI functions for Research & Development on the Duolingo English Test.As an expert software engineer and AI researcher who has also taught and studied abroad in two foreign countries, he is passionate about the applications 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.

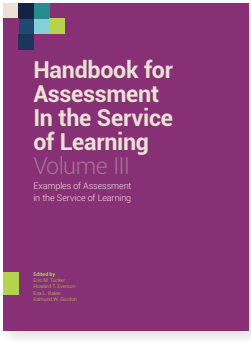
Jessica W. Younger, Ph.D., is an educational neuroscientist dedicated to developing effective interventions that empower learners to reach their full potential. With over a decade of experience, her work explores how individual differences shape learning, leveraging advanced statistical modeling and 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 II of the *Handbook for Assessment in the Service of Learning* moves from foundational principles to the conceptual tools and methods needed to build assessment systems that actively improve, not just measure, learning. Section I offers frameworks for learner-centered assessment—foregrounding formative practice, self-regulated learning, personalization and equity, validity, and social justice—so that technical quality and justice are co-equal design imperatives. Section II translates these ideas into practice: game-based learning, educative portfolios, dynamic learning maps, culturally and linguistically responsive co-design, redesigned score reporting, and analyses of learner–system interactions that turn digital traces into actionable evidence. Volume II of this *Handbook for Assessment in the Service of Learning* provides blueprints and validation guidance to inform assessment systems that support learners—bridging Volume I's foundations to Volume III's examples. It also advanced the series' proposition that assessment, teaching, and learning are inseparable.