

Formative Assessment in a Digital Learning Platform

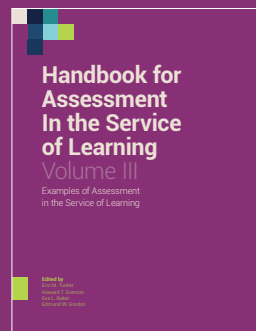
Kristen DiCerbo

UMassAmherst

University Libraries

Series Editors:

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





© 2025 by Kristen DiCerbo

The Open Access version of this chapter is licensed under a Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License (CC-BY-NC-ND 4.0).

ISBN: 978-1-945764-33-2

Suggested Citation:

DiCerbo, K. (2025). Formative assessment in a digital learning platform. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning*. University of Massachusetts Amherst Libraries.

Formative Assessment in a Digital Learning Platform

Kristen DiCerbo

Khan Academy

Abstract

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

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

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

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

Formative Assessment on a Digital Learning Platform

We live in a world where students engage in practice on digital learning platforms and are able to receive both necessary scaffolding to build their skills and immediate feedback to course correct. In addition, these platforms are able to collect and store information about these practice experiences, both the correctness of student responses and information about scaffolds used, attempts taken, and time spent. In distilling this information, digital learning platforms such as Khan Academy provide meaningful summaries about what students know and can do at a level of granularity that can help inform instructional decisions, essentially becoming formative assessment¹ tools.

Time Versus Information

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

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

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

¹ Formative assessment is the range of formal and informal procedures used in classrooms to help teachers and students understand learning while it is in process in order to adjust teaching and learning strategies.

It stands in contrast to summative assessment, which is conducted at the end of a segment of learning in order to understand whether a learner has achieved the intended outcomes.

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

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

A Brief Overview of Assessment at Khan Academy

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

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

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

Figure 1.

Equations with variables on both sides

CCSS.Math: 8.EE.C.7.b, 8.EE.C.7  Google Classroom  Microsoft Teams

Solve for f .

$$-f + 2 + 4f = 8 - 3f$$

$$f = \boxed{}$$

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

Mastery Learning

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

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

Setting Expectations for Progression

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

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

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

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

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

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

Supporting Motivation and Engagement

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

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

Feedback

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

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

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

Figure 2.

Which statement about genes is true?

Choose 1 answer:

A Genes are made up of DNA nucleotides called A, C, L, and S.

B Genes are part of larger structures called chromosomes.

INCORRECT (SELECTED)

Having different genes usually causes organisms to have the same traits.

Having different genes causes organisms to make different proteins, and to have different traits.

Related content



Genes, proteins, and traits



Genes, proteins, and traits

Figure 3.

Solve for m .

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

$$m = \boxed{}$$

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

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

$$4m + 3 = 15 - 2m \quad \text{Combine like terms.}$$

$$3 + 4m + 2m = 15 - 2m + 2m \quad \text{Add } 2m \text{ to each side.}$$

$$6m + 3 = 15 \quad \text{Combine like terms.}$$

$$6m + 3 - 3 = 15 - 3 \quad \text{Subtract 3 from each side}$$

$$6m = 12 \quad \text{Combine like terms.}$$

$$\frac{6m}{6} = \frac{12}{6} \quad \text{Divide each side by 6.}$$

$$m = 2 \quad \text{Simplify.}$$

3 / 3 The answer:

$$m = 2 \quad \text{Let's check our work! } \checkmark$$

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

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

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

Figure 4.



Figure 5.

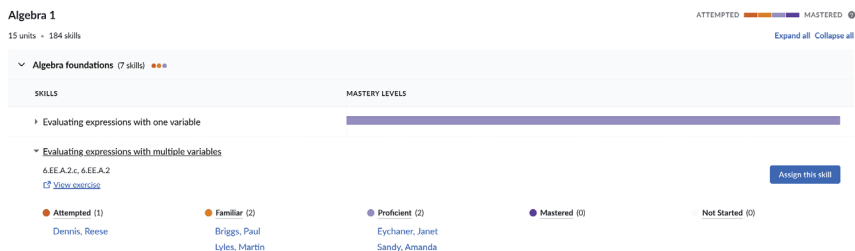
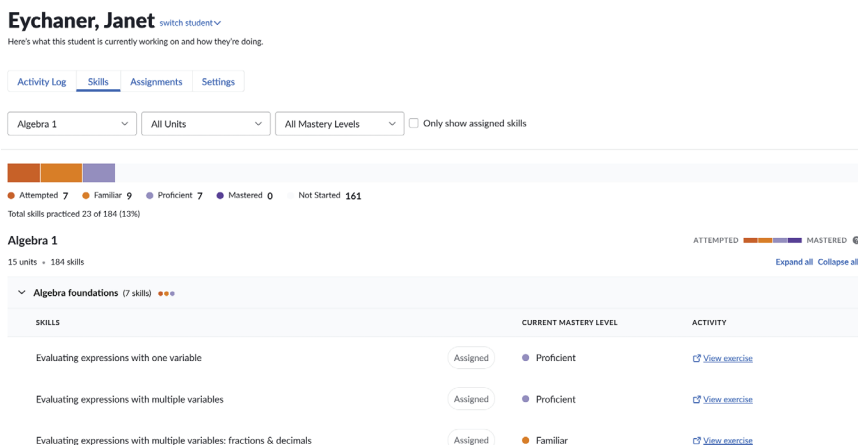


Figure 6.



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

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

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

Does it Work?

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

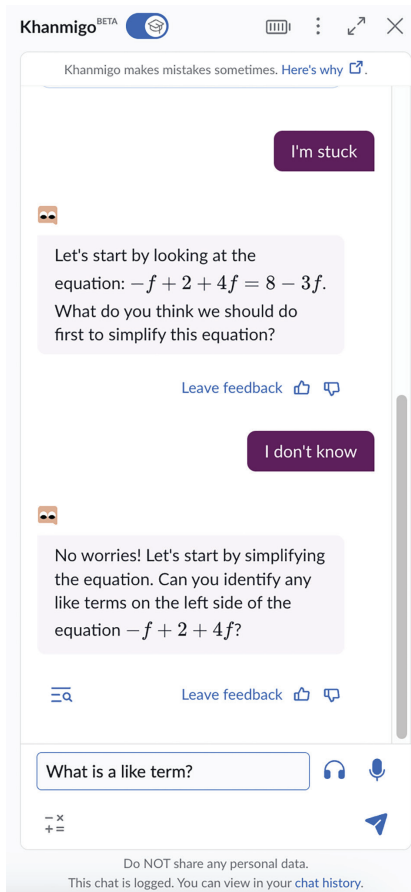
From an assessment perspective, the question of “does it work?” refers to validity, reliability, and fairness. Classroom formative assessments, especially teacher-made assessments, do not often undergo the close analysis of these factors to the level of rigor of summative assessments. The consequences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

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

Generative AI and the Future of Assessment

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

Khanmigo takes the power of the generative AI model and designs a specific education experience with it. For example, in math tutoring on Khan Academy, when a student wants to use the Khanmigo tutor, the student's input gets sent to the model along with instructions on how to act like a tutor. These instructions are based on research on what makes a good human tutor, which has been conducted over the past decades of trying to develop intelligent tutoring systems (e.g., Graesser, Person, Magliano, 1995). As a result Khanmigo, makes goal statements, course corrects when students are headed down the wrong path, and makes similar tutor moves that help the student get to the answer themselves (See Figure 7). In a writing coach application, separate instructions are sent to the model so it can evaluate various aspects of student writing (e.g., the ability of the introduction to capture attention, the use of evidence in an argument, etc.), provide that to students, and then engage in conversation about how to improve.



At the time this chapter is being written, there is significant excitement about the potential for generative AI to impact education, but it is early days in its development and it is not clear how much of the promise will be realized. In particular, education has typically moved slowly to adopt new innovations (Reich, 2020), often for good reason; the education of children is important and should not be subject to every fad that arrives on the scene. Ideally, evidence would be gathered on new interventions before they are scaled widely. Unfortunately, conducting the kinds of rigorous research done in, for example, the medical field

is also fraught with roadblocks and difficulty. As a result of numerous constraints, many decisions about which educational technology products to use are often based on word of mouth among district administrators and less information is gathered about effectiveness than would be desirable (Morrison, Ross, & Cheung, 2019).

In the current situation with generative AI, in the 2023–24 school year 53 school districts participated in a pilot of Khanmigo. Many did so with specific schools, domains, and/or grades in an effort to try out the tool before bringing it to scale. The state of Indiana released a request for proposals that allowed districts to receive funding for such pilots and also then ran teacher surveys to gauge their perceived usefulness (Appleton, 2024). The uses of Khanmigo clearly fell in the learning space, but give us some direction of how generative AI might impact the future of assessment.

Skills to Assess

Evidence-centered design (Mislevy, Almond, & Lukas, 2003) defines the domain model as the set of knowledge, skills, and attributes to be assessed. It is possible that the advent of generative AI opens up a new set of skills that should be assessed. In the workforce, many professionals are already using generative AI as an assistant. Software engineers at Khan Academy use the GitHub copilot to code with them. The engineer indicates what code they want to write, the AI copilot drafts code, and the engineer reviews and revises it. Similarly, many people who need to write text ranging from marketing emails to job descriptions are using the AI technology to create first drafts. The number and type of uses of generative AI suggests the importance of the skills of evaluation (of code and text) and editing is going to increase. However, evaluation and editing are rarely assessed currently, but should be considered in assessment research and development spaces.

New Task Models

The task model is the abstraction of the activity with which the person being assessed engages. Historically, the activity types used for assessment have been limited to what the technology available could support for large-scale automated assessment. When the only option available was optical scanners, multiple choice questions provided the best way to score a significant number of assessments quickly. As technology has progressed, variants, often called “technology enhanced items” appeared, including drag and drop, match and order, and more recently graphing, hot spot, and audio and voice items.

There has been a significant amount of work done on simulation and game-based assessment (e.g., Gobert & Sao Pedro, 2016; Shute & Ventura, 2013; Baker, Dickieson, Wulfbeck, & O'Neil, 2017). Both offer the possibility of more authentic tasks for students. The premise of these assessments is that rather than ask students a question about how they would do something, we can ask them to do that thing in a simulated environment. Ideally, the use of tasks like those in the real world should shorten the assessment's inferential distance (Behrens, DiCerbo, & Foltz, 2019), the theoretical distance between what we observe someone do and what we infer about what they know and can do from that evidence. For example, there is a relatively large distance between observing which option was selected in a multiple choice question about computer networking and inferring that someone can configure a network. By offering students a simulation, we can observe them engage in many of the actual tasks we are interested in assessing (Behrens, Collison, & DeMark, 2008).

There are downsides to the use of more authentic types of assessment tasks. First, the assessment time-to-evidence ratio can be high. In games, it is often the case that students engage in 30–45 minutes of gameplay and only generate a few pieces of evidence that provide information about the construct of interest. SimCityEDU, a game-based assessment of systems thinking in which students worked to diagnose why city residents were sick and fix the problem, demonstrated the trade-offs between game play and evidence. Ultimately, the problem students needed to diagnose was air pollution; the solution was adding new energy types and removing some of the coal-burning power plants. In terms of systems thinking, evidence consisted of things like placing in new energy types before removing the coal plants vs. removing the coal plants first (which would leave the city with no power, an indication of poor systems thinking). However, it took a significant amount of game play to diagnose the problem causing the residents' illness and then to uncover solutions. After extensive gameplay, only a few pieces of evidence ended up in the statistical models estimating systems thinking proficiency (Castellano et al., 2014).

SimCityEDU also highlighted that inferential distance remains in many simulations and games. In analyzing moves students made in the game, it became apparent that about 5% of students were bulldozing the entire city. What should we infer from this behavior in regard to systems thinking? Were they thinking about rebuilding the city from the ground up based on ways to eliminate air pollution? Was bulldozing

the city the ultimate in systems thinking? As we found out when we asked a handful of students, it was mostly because bulldozing is fun. We find with simulations and games, in many cases, to eliminate inferential distance, we need to use language and ask students what they are thinking.

The need for new functionality is especially apparent when the skill to be assessed is rooted in language, such as collaborative problem solving. PISA (Program for International Student Assessment) undertook an assessment of collaborative problem solving skill in 2015, including producing a detailed framework describing the skill (Foster & Piacentini, 2023; OECD, 2017). The team wanted to observe students actually engaged in collaborative problem solving. Doing that with other human learners though was technically difficult and introduced a considerable amount of variability. So, the assessment had learners interact with automated agents. However, due to the difficulty of processing and scoring natural responses, students were given multiple choice options from which to choose a response, rather than entering a free-form response. If the students were able to type anything they wanted in a response, there was no good system by which their automated collaborators could engage in conversation about the wide range of things the students might say. Drafting these dialogue trees was a large task even in the agent-based solution. The assessment led to informative results. However, the difficulty in managing language continued to result in a gap between what was observed and the inference to be made.

Over the decades, significant work has been done on automated essay scoring (Shermis & Burstein, 2013). Work that began with the identification of features that correlated to human scoring, such as essay length, matured into models that used the meaning of words to evaluate essays. Today, many programs score essays at the same level of agreement to humans that other humans do, not perfectly because humans also disagree, but at a high level. The difficulty with these programs is that they usually require training the model on the specific essay to be scored using hundreds or thousands of human-scored examples. Additionally, from a learning perspective, just getting a score is not sufficient to help a learner know how to improve.

Enter generative AI. It cannot solve all of the problems, but it can, in combination with solutions we already have, improve our existing assessments. First, the models, with proper prompting can engage in conversation. Those skills involving

dialogue can potentially be assessed directly rather than through selected responses. Additionally, instead of trying to infer why a student did something a certain way, whether answering a traditional math problem or bulldozing cities, the AI can ask them and engage in dialogue about what is being observed. The inferential distance can be further closed.

Generative AI can also be good at giving feedback on writing. In classrooms currently, students do not often get assigned longer essays because they require a large effort to grade. As a result of that effort, feedback is often delivered many days or weeks after the writing was done and little feedback likely influences performance on future writing assignments. Khan Academy now has a writing coach feature that walks students through writing assignments and then provides feedback on aspects of the students' drafts. For persuasive essays, for example, Khanmigo gives feedback on students' introduction, use of evidence, structure, conclusion, and tone and style. The ability to provide feedback specific to elements of essay writing won't happen "out of the box" with a large language model, but it can with applications specifically designed to use the models for education. To get Khanmigo to provide this feedback, we split each area of feedback into a separate prompt. Each of those prompts contains instructions based on what writing teachers look for in that element, telling the model what to look for. Designers and engineers then created the means by which students could edit and Khanmigo could "see" the changes that students are making and converse with the students about them. The feedback functionality of generative AI has the potential to fill in the feedback gap in most automated scoring of writing.

Finally, generative AI has the potential to allow for more individualization of activities in ways that will enable for different background knowledge and experience to be considered. Even with the computer-adaptive test, the adaptivity focuses on the students' measured achievement level and the difficulty of the items. All students are working from the same item pool. However, we know that questions can be differentially difficult depending on familiarity with the (sometimes irrelevant) context of the question. The classic example of the impact of background knowledge on comprehension used for those familiar with the majority American culture is to give a reading passage about baseball and one about cricket. Americans with a deeper knowledge of baseball and nearly no knowledge of cricket do much worse on reading comprehension questions about cricket.

It is possible that generative AI could be used to adjust the background knowledge of questions for students in ways that do not penalize some students for their lack of familiarity with contexts. The adjustment of context in an assessment question has been impossible because it did not make sense to assume familiarity with given contexts based on someone's rough demographic profile. Now, it is possible that AI solutions could be used to tackle the problem of personalization. At Khan Academy, students can choose to converse with Khanmigo about their interests. Khanmigo probes on different topics, from food to sports to hobbies, and records up to 10 interests in the student's profile. Students can always go in and modify or delete what they have entered. These interests are then injected into different prompts to guide Khanmigo so the conversations can incorporate the interests. Currently, Khanmigo can adapt questions during a conversation to incorporate these interests. Still, the responses to the adapted question do not feed back into the mastery system, largely because we have not built the infrastructure for information from conversations to be incorporated into scoring and mastery mechanics. That said, there is a clear research and development need for mechanisms by which to equate items with differing contexts, potentially created in the moment of administration.

Reporting

As mentioned above, many teachers do not make use of data from digital systems. Despite valiant efforts at design research (Zapata-Rivera, 2018; Zenisky & Hambleton, 2015), score reports primarily do what their name suggests, and report scores. Generative AI offers the potential to let consumers of assessment results have conversations about the results, including asking questions about what they mean and getting recommendations from them. At Khan Academy we now have an AI tool for teachers called Class Snapshot where Khanmigo first gives a summary of student performance in the class, including the time spent and skills leveled up. The statistical summary is done with a calculator and fed to the large language model in order to ensure mathematical accuracy. The teacher can then ask questions such as "who needs help adding fractions?," "who should I group together for a lesson on multiplying decimals?" and "what should I assign to my students next for practice?" The latter will produce groups of students of similar skills and suggest Khan Academy content. The teacher can then interrogate the model's responses and make decisions about whether to accept the recommendations, allowing teachers to obtain, not just data given to them, but clear options for action based on that data.

Challenges to Psychometrics

As the field starts looking at assessment in technology rich environments, some of the existing rules and procedures may need to be revisited (DiCerbo, Shute, & Kim, 2017). Many of the techniques used for measuring the psychometric properties of assessment were developed in the context of standardized assessment, consisting of discrete items specifically designed to assess a single construct, scored as correct or incorrect. Much of the evidence gathered from assessment in technology-rich environments (e.g., time spent, sequence of events) is not scored as correct/incorrect and often relates to multiple constructs. In addition, there is often variability in what activity is presented to different learners depending on their own progress and choices.

As described above, currently Khan Academy proficiency is a strong predictor of external assessment performance. As new methods of assessment are developed, the standards for acceptable correlation levels with external measures are not clear. For example, if a new generative AI-based item type purports to be a better measure of a construct because it eliminates unknown background context, we might expect a lower correlation to existing measures. An open question for discussion in the field then becomes: how do we demonstrate an innovative assessment is actually a better measure of a construct than an existing assessment?

The potential lower correlations will also present a challenge to adoption of new forms of formative assessment as long as schools and districts are held accountable through their scores on traditional assessments. Decision-makers in schools will want assessments that predict whether students are on track to be successful on end-of-year assessments even if that end of year assessment is less perfect.

Do We Need Summative Assessments?

Given the relatively large amounts of data about student performance coming from interactions with digital learning environments, some have asked whether we need summative assessments. In fact, John Behrens and I have laid out a vision for the future in the “digital ocean” where, because we have so much data from daily learning interactions, we do not need to ask people to stop and take a test (DiCerbo & Behrens, 2014). However, we are not at that place at the moment. The data collected by Khan Academy is vast; there is a large data lake full of student

interaction data. However, there is a lot of noise in the data. Students start working on a problem and walk away, then return and skip to other exercises on other skills. Student choice and agency was built into the platform on purpose to allow students to pursue individual interests. Students may be working together to solve problems with their peers (which is acceptable in a learning context). In a recent classroom visit, students were observed working in pairs and using one student's Khanmigo account to ask questions when they got stuck. Context information of this kind is not gathered on the platform and while it could be, inserting points of friction in the experience, for example, requiring students to enter names of those they are working with, decreases the likelihood of actually engaging in the learning activities. Students begin conversations with Khanmigo but then drop off, maybe because they get the help they need but perhaps because they didn't get the help they needed. For purposes of formative assessment, where the decisions being influenced are around what should be taught the next day, and there are teachers and parents in the loop to make adjustments if what is indicated by the assessment is a little off, this noise is acceptable. However, if more consequential decisions were to be made, with less chance of correcting for error, these measures are likely too unreliable in their current state to be fit for that purpose.

Concluding Thoughts

The use of generative AI to solve some of the long-standing problems in assessment sounds quite promising (or perhaps quite daunting) and there is great potential, but there is also much research to be done before these models can be used in higher stakes assessment. Even for formative assessment, a big challenge comes from the fact that large language models are, by definition, probabilistic. The responses the model gives, even from the same instruction and student input, vary each time the model produces a new response, which impacts standardization, but it could also impact the extent to which the model prompts students for more information or gives help or hints. Models can do well nearly all the time but occasionally give an odd response. In low-stakes environments, with a teacher available, wrong or illogical responses can be addressed but it would be a significant concern in higher stakes situations. More work is needed before generative AI-based tasks or scoring can be validly and reliably used for high stakes decisions.

More generally, existing digital learning experiences offer learners the possibilities of nearly unlimited practice with immediate feedback. The amount of information gathered from these opportunities is sufficient to inform instructional decisions about what students need support on, where they are succeeding, and what they should work on next. The systems are ideal for setting expectations for what is to be learned over time and providing students with feedback in ways that support learning. The introduction of generative AI offers the ability to improve on the use of information from assessments to inform instruction and also to build equitable experiences where students are not penalized for construct-irrelevant differences in background knowledge. Much of the ability to provide instructionally relevant information comes from the fact that these data are gathered over time, providing the ability to capture multiple instances of students solving problems at the skill level. At the same time, information gathered during informal practice also results in significant noise in the data, which cautions against its use in high-stakes decisions. Ultimately, data from student experiences on one platform will never capture the sum of all they know and can do, but it can help give us more information about students at a more granular level if used with care.

References

- Appleton, A. (2024, June 17). Indiana schools embrace AI, but seek to 'keep humans in the loop' *Chalkbeat*. <https://www.chalkbeat.org/indiana/2024/06/17/students-use-ai-pilot-programs-in-class/>
- Baker, E., Dickieson, J., Wulfbeck, W., & O'Neil, H. F. (Eds.). (2017). *Assessment of problem solving using simulations*. Routledge. <https://doi.org/10.4324/9781315096773>
- Behrens, J. T., Collison, T. A., & DeMark, S. (2008). The seven C's of comprehensive online assessment: Lessons learned from 36 million classroom assessments in the Cisco Networking Academy Program. In L. A. Tomei (Ed.), *Online and distance learning: Concepts, methodologies, tools, and applications* (pp. 2578–2592). IGI Global. <https://doi.org/10.4018/978-1-59904-935-9>
- Behrens, J. T., DiCerbo, K. E., & Foltz, P. (2019). Assessment of complex performances in digital environments. *Annals of the American Academy of Political and Social Science*, 683, 217–232. <https://doi.org/10.1177/0002716219846850>
- Castellano, K., Hoffman, E., Bauer, M., Bertling, M., Kitchen, C., Jackson, T., Oranje, A., DiCerbo, K., & Corrigan, S. (2015). *Game-Based Formative Assessment for Argumentation: Mars Generation One: Argubot Academy* [Conference presentation]. Annual meeting of the American Educational Research Association, Chicago, IL.
- Choi, J., DiCerbo, K., Ventura, M., Lai, E., Wood, J., & Iverson, J. (2019). *Measuring proficiency using interactive simulation data: Empirical comparison of evidence aggregation methods* [Conference presentation]. National Council on Measurement in Education Annual Meeting, Toronto, Ontario, Canada.
- DiCerbo, K. E., & Behrens, J. T. (2014). *The impact of the digital ocean on education* [White paper]. Pearson. <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/DigitalOcean.pdf>

- DiCerbo, K. E., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations of ongoing assessment. In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy*. Springer (pp. 1–21). https://doi.org/10.1007/978-3-319-17727-4_66-1
- Foster, N., and M. Piacentini (Eds.). (2023). *Innovating assessments to measure and support complex skills*. OECD Publishing. <https://doi.org/10.1787/e5f3e341-en>.
- Gobert, J. D., & Sao Pedro, M. A. (2016). Digital assessment environments for scientific inquiry practices. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 508–534). Wiley. <https://doi.org/10.1002/9781118956588.ch21>
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495–522. <https://doi.org/10.1002/acp.2350090604>
- Grimaldi, P. (2023, November 16). Multiple studies show Khan Academy drives learning gains: Evidence for our platform's effectiveness. *Khan Academy Blog* <https://blog.khanacademy.org/multiple-studies-show-khan-academy-drives-learning-gains-evidence-for-our-platforms-effectiveness/>
- Guskey, T. R. (2022). *Implementing mastery learning*. Corwin Press.
- Kulik, C. L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265–299. <https://doi.org/10.2307/1170612>
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29. <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>
- Morrison J. R., Ross S. M., Cheung A. C. (2019). From the market to the classroom: How ed-tech products are procured by school districts interacting with vendors. *Educational Technology Research and Development, 67*(2), 389–421. <https://doi.org/10.1007/s11423-019-09649-4>

- OECD (2017). *PISA 2015 results (Volume V): Collaborative problem solving*, PISA. OECD Publishing. <http://dx.doi.org/10.1787/9789264285521-en>
- Oreopoulos, P., Gibbs, C., Jensen, M., & Price, J. (2024). *Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence* (No. w32388). National Bureau of Economic Research. <https://doi.org/10.3386/w32388>
- Reich, J. (2020). *Failure to disrupt—Why technology alone can't transform education*. Harvard University Press. <https://doi.org/10.2307/j.ctv322v4cp>
- Senko, C. (2019). When do mastery and performance goals facilitate academic achievement?. *Contemporary Educational Psychology*, 59, article 101795. <https://doi.org/10.1016/j.cedpsych.2019.101795>
- Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge. <https://doi.org/10.4324/9780203122761>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press. <https://doi.org/10.7551/mitpress/9589.001.0001>
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. <https://doi.org/10.4324/9781410605931>
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 487662. <https://doi.org/10.3389/fpsyg.2019.03087>
- Yamkovenko, B. (2023, September 25). Why Khan Academy will be using "skills to proficient" to measure learning outcomes (and you should too!). *Khan Academy Blog*. <https://blog.khanacademy.org/why-khan-academy-will-be-using-skills-to-proficient-to-measure-learning-outcomes/>

Zapata-Rivera, D. (Ed.), (2019). *Score reporting research and applications*. Routledge.
<https://doi.org/10.4324/9781351136501>

Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 585–602). Routledge.
<https://doi.org/10.4324/9780203102961>