Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2020, Vol. 39, No. 3, pp. 90–93

Assessment for Learning with Diverse Learners in a Digital World

Kristen DiCerbo, *Khan Academy, Palo Alto*

Abstract: *We have the ability to capture data from students' interactions with digital environments as they engage in learning activity. This provides the potential for a reimagining of assessment to one in which assessment become part of our natural education activity and can be used to support learning. These new data allow us to more closely examine the processes learners use to solve problems, not just their final solutions, so educators can be more targeted in their intervention. New capabilities may address some of the consistent criticisms of assessment, but they also present new dangers, particularly for diverse learners. Systems are in danger of replicating the biases from our non-digital world in our digital tasks and in the creation of the scoring tools that accompany them.*

Keywords: algorithm, artificial intelligence, assessment for learning, digital ocean, machine learning

The primary goal of educational assessment is to infer attributes about a learner from observation of their performance on a given activity. Historically, this has been accomplished by presenting students with a defined set of items, usually framed as questions, to which they respond. We call these tests. However, we now live in a digital world where students engage in a variety of rich activities from which we can gather detailed data. This ability to capture and analyze information from students' daily learning activity should fundamentally change how we think about assessment (DiCerbo & Behrens, 2014). This (relatively) new ability allows us to potentially make inferences about what learners know and can do as they are engaged in learning activity, resulting a blurring between learning and assessment. However, this potential sea change in the relationship between learning and assessment must also include new ways to think about assessment that improve our inferences about all learners. This article discusses some of the advances in assessment in the digital world and then describes some potential areas of focus to ensure new practices in assessment design support the assessment of diverse learners.

Assessment in the Digital Ocean

Historically, in order to observe and evaluate learner behavior at scale, for example in order to make school accountability judgments or decisions about admission to college, it was necessary to employ fixed-time, forced-response tasks that were distal proxies of day-to-day activities. Specific data needed to be collected in a specific amount of time, and therefore required efficiently implemented, forced tasks.

In contrast, in the educational practice of teachers, the assessment model has been much more fluid: interactions occur repeatedly over time, the context of activity is consistent with current goals, observations and performance are made and characterized formally and informally, and the subsequent activities of instruction are introduced in light of previous performance. In such contexts, assessment becomes a natural ubiquitous and unobtrusive (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2007) side effect in the natural environment or what Shute calls stealth assessment (Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

The first paradigm benefits from formalization of rules of evidence, standardization of experience (allowing for comparability), collection of data, and data analytics while suffering from decontextualized tasks. It also tends to emphasize one correct answer and one right path to get there. The second paradigm benefits from contextualized and "authentic" tasks while suffering from informal data collection, rules of evidence and analytics, and difficulty with comparability and communication outside a specific classroom. The emerging universality of digital tasks and contexts provides the opportunity to unify the strengths of each of these approaches. As the digital instrumentation needed for educational assessment increasingly becomes part of our natural educational activity, the need for intrusive assessment practices that conflict with learning activities diminishes.

This vision, described in Shute, Leighton, Jang, and Chu (2016), does not involve administering assessments more frequently (e.g., each week, each day) but, rather, continually collecting data as students interact with digital environments. It relies on what Shute (2011) calls stealth assessment, the process by which data are unobtrusively gathered while students are playing/learning in carefully designed environments that then allow inferences to be made about relevant competencies. As the various data streams coalesce, we obtain more evidence about what students know and can do across multiple contexts. The vision of assessment in technology-rich environments involves high-quality, ongoing, unobtrusive assessments that can be aggregated to inform a student's evolving competency levels (at various grain sizes) and also aggregated across students to inform higher-level decisions (e.g., from student to class to school to district to state to country).

However, even without arriving at this vision of ongoing assessment, there are other levels of digital assessment implementation that take advantage of the possibilities the digital world has to offer. DiCerbo and Behrens (2012) layout a categorization scheme for thinking about these possibilities. Level 1 merely replicates traditional activities in a digital environment. Level 2 involves new assessment activities that allow new types of performances and new types of information. This includes, for example, complex simulation tasks that require integration of multiple skills and collects data about the process learners use to solve the problem. Level 3 involves embedding assessment in the learning environment, for example as a game that is threaded through semester-long courseware that serves as both a learning tool and a means to understand what learners know and can do. Level 4 is the vision described above of a system that integrates inputs from multiple digital activities over time to build an understanding of what learners know and can do. Even at level 2, there is potential progress being made toward not just "assessment for learning" but learning as assessment.

Process Data

One objection to traditional, standardized measures of assessment is that they produce a single, binary data point from each question: correctness or incorrectness of the final answer. This does not help us understand how a learner arrived at a particular response, which can be key in remediation (and also key in identifying whether there are potentially other correct answers). Digital environments allow us to capture many features of a learner's performance on an activity, for example time, sequence of activity, and counts of actions. The game SimCityEDU is designed to assess systems thinking, particularly the ability to understand how actions in a system have multiple, interrelated effects. Players are dropped into a city that is already built and is experiencing problems where they are asked to diagnose and fix the issues. It turns out that there is significant air pollution in the city, and that coal plants are causing a lot of it, however, coal plants also provide power to the city. By observing a sequence of actions, whether the player installs alternative forms of energy prior to bulldozing the coal plants or not, provides us evidence about whether the learner understands that the coal plants have multiple effects on the city, as opposed to just the single effect of air pollution. Gathering multiple observations of similar choices, allows us to build models of their levels of systems thinking (DiCerbo, Castellano, Jia, Mislevy, & Jin, 2015).

This ability to capture log file data opens the possibilities of providing insight into a learner's process of problem solving, not just their final answer. This insight can then help point to particular areas where a learner's solution might differ from the "prescribed" solution. Over time, and thousands or tens of thousands of users, we will be able to identify even less common solution paths in digital experiences, often with the help of exploratory data analysis (Behrens, DiCerbo, Yel, & Levy, 2012; DiCerbo et al., 2015). In the SimCityEDU game, over time it became apparent that a small percentage of players "solved" the air pollution problem by bulldozing the entire city. Rather than simply tagging these as "wrong," it was instructive instead to explore why learners follow a particular path. In the game, interviews with learners revealed that bulldozing things is fun, that's why players chose it as a solution, and we should not infer anything about the systems thinking ability, positively or negatively, for those who bulldozed the whole city.

In the classroom, conversation around why a learner took a particular path of action can open up a rich dialogue about how a learner has framed a problem and selected a solution option. Clearly, this kind of discussion is beyond the means of our digital capabilities; it is the domain of a capable teacher. Ideally, the digital experience provides a gateway or launching point for these conversations that go far beyond discussions of correctness. Teachers have been exhorting students to "show your work" for decades and the digital world can help make that a default for all students, and help uncover patterns in the process.

Making Sense of Complex Data

Learner-generated data in new digital environments can be complex. The new data can include written, verbal, and click stream input, as well as the construction of many types of media (e.g., video) by the learner. They may be produced by one or more learners who are interacting with each other and they may be highly dependent on the number of states or contexts within the digital environment or simulation. Making inferences of performance in complex environments is often difficult even for instructors (e.g., human raters), who suffer from fatigue, rater drift, and known biases like the halo effect. Early attempts at computer scoring of digital performances used simple rule-based approaches. In the SimCityEDU example above, this included making a binary rule that scores whether a learner builds an alternative energy source prior to bulldozing the coal plant. However, these approaches quickly become overwhelmed by the near-infinite potential branching points, open-ended responses (e.g., written or spoken responses), and complicated interactions of different types of data (e.g., a mixture of student behaviors or actions in the digital environment with responses).

To address this explosion of complexity and data, advances in machine learning (ML)—and artificial intelligence (AI)—based analyses of data have been developed to support immediate processing and automated inferences of student performances. The typical development process of an ML-based approach requires researchers to (Behrens, DiCerbo, & Foltz, 2019):

1. Define the constructs of interest.
2. Extract features of the performance in the digital environment to match those constructs. The features are the elements described above, such as time, counts and sequence of activity, and final state of the solution. When a digital environment can capture every mouse click, it can be difficult to know what is important. The process can use humans, who may have designed the experience specifically to elicit certain evidence, computers that seek to identify features related to participants' level of knowledge or skill, or a combination of both (Sao Pedro).
3. Identify a "ground truth" about performance. This is typically expert scores, ratings, or comments about the performance.
4. Train the ML model to associate features to characterizations of expert judgments. ML will learn to associate the selected features with the expert judgments.

5. Validate the performance of the models on new data from new populations. While ML is good at predicting performance on a particular training set, it is critical that the assessment model is generalizable across the intended population.
6. Build filters to detect unusual input. With the much wider range of responses or performance patterns, it is critical to be able to detect when the performance pattern differs greatly from what the ML model has been trained on. For example, in a problem-solving task, the system should detect if a user is using a highly different pattern of actions to solve the problem. These can be flagged for instructor review and/or used for updating/retraining the model.
7. Finally, once the automated assessment has been developed and validated, it can be implemented in training and assessment systems. Despite the assessment component being automatized, human oversight should remain part of the process to ensure continued reliability and validity.

Dangers Lurking in the Digital Ocean

Assessing complex performance behavior requires more than having detailed data about the learner (Behrens & DiCerbo, 2014). It also means being able to make intelligent inferences about that performance. The fundamental failure of assessment occurs when we make incorrect inferences about learners. Assessment in the digital world offers promises of better inference through understanding of process, but there are places where the new possibilities for assessment create new ways to make incorrect inferences, particularly for diverse learners. It is also possible that these new approaches will merely replicate many of our existing systemic biases and inequalities. There are at least three areas in assessment design where we should consider these possibilities: task design, evidence identification, and evidence accumulation.

Task Design

Digital assessment activities are touted for being more authentic, bringing in more context, and being more similar to the real-world scenarios into which we want learners to generalize and apply their skills. However, this raises the question of whose context the digital scenarios are modeled after. Including more *unfamiliar* context potentially introduces construct-irrelevant variance into a performance in unequal ways. Students who are more familiar with the context will exhibit better performance than those less familiar, for reasons having nothing to do with their target knowledge and skill proficiency. This has always been a potential problem in things like simple math word problems and reading comprehension, but creating more immersive, developed scenarios in a digital environment has the potential to greatly exacerbate the problem. The digital environment likely contains far more extraneous information that an unfamiliar student will have to sort through without clues as to what is important and what is not. Even if the scenario is legitimately one to which we want skills to generalize, it adds another layer of the complexity to making inferences about the skills of interest.

Evidence Identification

As described above, one step in digital assessment is the identification of which features of performance constitute evidence. Here again, whether a human or a machine makes decisions, there are opportunities for bias. People have their own understandings about the elements of a performance that make it poor, fair, or good. In some cases, these understandings stem from a particular context and world view, and are blind to other responses that are legitimately appropriate from another world view. As an example, in judging the performance of someone configuring a computer network, an assessment expert might posit that high use of help commands is associated with more novice performance. Instead, data from actual programming experts shows they are frequent users of help because their expertise comes not from memorizing all the commands but from knowing the approach to take to do the configuration.

Relying on the data to tell us the important features is also not the answer because the data available themselves often do not contain sufficient numbers of people from diverse backgrounds to capture the potential variability in what could be important features. All data methods at some point come back to training a model using human judgment, for example, a human identifying which performances overall are indicative of novice versus expert, or scored in various rubric categories. Other approaches have humans suggest a large pool of features and then have the machine learning techniques narrow down that pool, or vice versa. In all these cases, the algorithms are finding patterns to match human judgment.

Evidence Aggregation

Once the relevant features of a performance are identified, we “train” a model how to combine these features into a score (in some methods, the evidence identification and aggregation all are part of the same procedure). Again, we are using human judgment as the “gold standard” by which to train our models when part of the argument is that human judgment is biased and often does not recognize the potential for what humans have determined to be the “wrong” answer to in fact be correct in some instances and contexts. We are replicating our existing systems with our technology in our new systems.

Second, we often are using flawed data with which to train the system. If the data does not contain adequate representation of diverse responses, the algorithms will not learn how those responses should be evaluated. The potential for problems can be seen in the classic example of Amazon’s attempt to build a resume-screening tool (Dastin, 2018). In the historical data, men had been the successful candidate far more than women. Women were not represented as successful candidates in the dataset. Although the developers did not include gender as a variable to be used in selection, they used resumes and hiring decisions from the past to train the model, and the model ended up factoring in proxies for gender, like attending a women’s college, to replicate those past decisions.

Finally, finding ways to cut out the human influence is not a viable solution either. Many algorithms developed are “black boxes” meaning it is very difficult to know and communicate how the model has combined features in complex ways to arrive at the answer. The algorithm is doing everything possible to maximize predictive accuracy, but that means it may be doing many things we don't want it to do, and we can't find out if it is. Instead, we have to test whether the outcomes of the algorithm are fair, which is difficult in itself, as evidenced by the lack of attention paid to consequential validity in the assessment literature, particularly on new forms of assessment (DiCerbo, Shute, & Kim, 2017).

Concluding Thoughts

Concerns with existing assessment systems are as follow: (1) they do not help inform classroom instruction, (2) they do not make accurate inferences about diverse learners, and (3) the things they ask learners to do are far removed from the real-life applications of knowledge and skill we desire them to be able to master. Digital environments offer the promise of addressing these concerns. They can place students in simulations of real-world environments that present authentic problems. Through collection of data during learning activity, they can break down the barrier between learning and assessment, where it is no longer a requirement to stop learning activity in order to take a test. Finally, the collection of process data offers the potential of understanding how students got to an answer and can serve as a catalyst for rich conversations about why learners took a particular solution path. This could help us improve our inferences about learners and particularly those whose solutions take them off the most commonly accepted path.

However, none of this promise is guaranteed and there is reason for concern about these new digital affordances. Overall, there is a serious risk that we replicate our existing biases in our new digital systems. First we will replicate the majority worlds in the tasks and contexts we model in the digital world. Then we will select the features or variables in performance that are important in our existing, flawed systems. Finally we will train the systems on data that already contain our known biases. There are actions that can be taken by all those involved in the learning and assessment research process to try to prevent this dystopian outcome. There is a small community who understands how evaluation of complex digital performances is done, and it is not highly diverse nor often focused on addressing issues of diversity. Everyone in the community needs more awareness and conversation around the issues of diversity in digital assessment. There is potential that our digital world can help address some of the persistent concerns about assessment to yield better outcomes for all learners, but there are major pitfalls that must be avoided along the path to a better future.

References

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.

Behrens, J. T., DiCerbo, K. E., Yel, N., & Levy, R. (2012). Exploratory data analysis. In I. B. Weiner, J. A. Schinka & W. F. Velicer (Eds.) *Handbook of psychology: Research methods in psychology*, 2nd ed. (pp. 34–70). New York: Wiley.

Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2007). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, and H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80) Mahwah, NJ: Routledge.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

DiCerbo, K., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, T. G. (2015). The role of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, and D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 319–42). New York, NY: Springer.

DiCerbo, K., & Behrens, J. (2012). From technology enhanced assessment to assessment enhanced technology. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, April 15.

DiCerbo, K. E., & Behrens, J. T.(2014). *Impacts of the digital ocean on education*. London: Pearson.

DiCerbo, K., Castellano, K. E., Jia, Y., Mislevy, R. J., & Lin, J. (2015). Analysis approaches for modified and original game based assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 15.

DiCerbo, K., Shute, V. J., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations. In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–21). New York, NY: Springer.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M-W. (2016). Advances in the science of assessment. *Educational Assessment*, *21*(1), 34–59.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.