



Invitational Research Symposium on
Technology Enhanced Assessments

Intended and Unintended Deceptions in the Use of Simulations

John T. Behrens, Kristen E. DiCerbo, and Steve Ferrara

May 7–8, 2012



Intended and Unintended Deceptions in the Use of Simulations

John T. Behrens, Kristen E. DiCerbo, and Steve Ferrara

Pearson

Executive Summary

Simulation designers set out to deceive learners into acting in natural and often complex ways on the assumption that this will provide relevant data to understand learners' knowledge, skills, or other attributes. In this paper, we discuss how this might be true in some ways but that, at other times, designers and analysts themselves may be deceived by a mismatch between simulation features and intended interpretations. To set the stage for applications of simulations to assessment contexts, we discuss the language of assessment delivery from the evidence-centered design (ECD) framework (Almond, Steinberg, & Mislevy, 2002) and compare this with traditional conceptualizations of assessment delivery that may limit full use of simulations. In addition, we provide a framework to discuss variations in the use of simulations in assessment interactions. Finally, we discuss specific simulation errors that undermine intended interpretations, including oversimulating, undersimulating, uniquely simulating, passively simulating, and simulating blindly, and the role that ECD can play in avoiding errors. Examples of different simulations and their use in various environments will illustrate the concepts.

Introduction

We believe that the technological change that we are currently engaged in as a society is transformational beyond our imagination and that new languages and concepts are required to tap the possibility (Behrens, Mislevy, DiCerbo, & Levy, 2012). Simulations, as a genre of interacting with machines, the environment, and other individuals, represent one of those areas of dramatic transformation that people are only now starting to explore 30 years into the personal computing revolution and not even 20 years into the communication network revolution. To unlock the power of new conceptualizations, we first discuss some key aspects of assessment delivery as described in the four-process language of evidence-centered design (ECD; Almond, Steinberg, & Mislevy, 2002). Next we review some of the common, perhaps hidden, assumptions commonly held about assessment structure. We do this to shed light on possible biases, in the hope everybody can improve their ability at taking alternative perspectives by understanding the hidden perspective they may already unknowingly hold. We then turn to an extension of the ECD language regarding constructs that may help us understand the role of simulation in the assessment development process. Subsequent sections review different aspects of simulation in light of the ECD delivery framework and the new language.

Four-Process Delivery Model of Evidence-Centered Design

Almond et al. (2002) described a four-process delivery model as part of the ECD framework. This aspect of the framework provides a highly abstracted language that describes a wide range of traditional assessment activities, as well as new forms of assessment including intelligent tutor systems (Almond et al., 2002), students working through open-ended simulation tasks (Frezzo, Behrens, & Mislevy, 2010; Williamson, Bauer, Steinberg, Mislevy, & Behrens, 2004), multistudent interactions in role-playing or simulated situations (Shute, 2011), and games (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008).

The key aspect of the framework is to deconstruct assessment delivery into four processes. As summarized in Table 1, the four processes are activity selection, presentation (of the activity), response processing (evidence identification), and summary scoring (evidence accumulation). The four stages of assessment delivery are presented in Table 1, along with specific examples from computer adaptive testing (CAT), in-class worksheet activity, and the use of electronic games.

Activity selection is concerned with choosing the next interaction with the learner. In a CAT setting, this typically means choosing an item that maximizes the test information function (cf. Wainer, 2000), although other approaches are available. In games, it may mean choosing activities that maximize motivation (Behrens et al., 2008) or, in dynamic tutors, maximizing change in the posterior distribution (Shute, Hansen, & Almond, 2008). Part of the allure of game-based environments is that the interactions are chosen rapidly and seamlessly so the end user has no sense of interruption or purposeful activity selection.

Table 1. Elements of the Four-Process Model With Examples

| Stage | Stage goal | Stage name(s) | CAT question example | Teacher in class example | Digital game example |
|-------|------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|-------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------------|
| 1 | Determine the next interaction | Activity Selection/ Task Selection | Pick next "item" | Decide what to do next | Pick next quest |
| 2 | Provide interaction; collect work product | Presentation | Provide next question; collect answer | Provide worksheet; collect when complete | Provide new quest, level, etc.; record all actions |
| 3 | Identify important features in the work product; create observations (tags of those identifications) | Response Processing or Response Scoring or Evidence Identification | Compare answer to key; determine correctness | Review and score answers; discuss variations | Score behavior; assign badges, tools, etc. |
| 4 | Combine, weight, and otherwise synthesize observations | Summary Scoring or Evidence Accumulation or Evidence Synthesis | Update estimated ability using IRT or similar | Sum points; update grade-book and teacher's mental model of student | Update overall score and resources for future activities |

Presentation is a central and unique idea to the four-process model. It is concerned with the interaction the learner has with the world and the output of that interaction, called work products. While a multiple-choice question is specifically constructed to constrain the work product to be a single (or multiple) choice from a list of possible choices, there are many forms of open response that can be used as work products as well. In some systems, multiple work products can be used simultaneously. For example, Frezzo, Behrens, DiCerbo, and Chen (2012) described a complex simulation-based system in which work products include the final state of a possibly large computer network, as well as the simulated results of specific networking functioning tests. They report additional research underway to incorporate log files as work products as well. Following Shute (2011), DiCerbo and Behrens (2012) argued that as daily activity becomes increasingly digitalized, the sources of presentation and work product for assessment will become increasingly integrated into the fabric of daily life. That is to say, the presentation and interaction processes that generate digital data for an assessment will be embedded in the systems of daily activity. This blurs the distinction between tests and naturalistic assessment, as well as formal and informal learning and inside/outside classroom distinctions (DiCerbo & Behrens, 2012).

Response processing (evidence identification) is concerned with the process of pattern recognition on the work product to make specific observations saved as observables. Observables are symbolic indicators of some meaningful observation that may be nominal, ordinal, or measured. Evidence identification is the process that translates the work of the learner into the low level observations of the expert (person or system). This process applies deterministic or probabilistic rules to

observe relevant aspects of the work product: you computed this as expected, your sentences are very long, you use rare words, you missed a step, you reference key ideas, your answer matches the key, and so on. Behrens et al. (2012) described how advances in data mining and other computational approaches will provide new insights and possibilities for information extraction from this process and thereby improve the inferential value of new forms of data.

One important implication of separating the presentation and response scoring processes is that assessments can be designed with varying combinations of presentation, work product, and response scoring. For example, a heterogeneous set of students can be provided with a single set of activities (write an essay) with differential response scoring rules applied depending on where they are in the developmental trajectory. This is, in fact, a common method for classroom teachers who regularly apply different rules of behavior to students based on their unique needs.

Summary scoring is concerned with the synthesis of information from the repeated generation of observables that happen across interactions (items, tasks, activities). Psychometric models undertake this synthesis using probabilistic models to reflect the nondeterministic relationship between latent constructs and the exhibition of specific human outcomes under specific circumstances (Mislevy, Behrens, DiCerbo, & Levy, in press).

Evidence-Centered Design and the Common Assumptions in the Item Paradigm

We use the term *item paradigm* to represent our impression of common assumptions that assessment designers and policy makers (and ourselves at different times) have, or have had, about the fundamental aspects of assessment. We think it is important to articulate these older assumptions so that the set of new possibilities can be contrasted properly with the old. The ideas that follow are not doctrinal assumptions of any one group or perspective, but rather our impression of common practice based on our collective work in a range of assessment environments.

While individual subscription to these tenets may vary, we see the following beliefs as limiting in comparison to what is possible in the new age of technology:

- Items consist of questions
- Items have answers
- Items measure correctness
- Items measure one thing

Items Consist of Questions

The ubiquity of the item as question is such that it is sometimes difficult to imagine another form. Whether the response format is open response or closed response, there is typically an assumption that the response is a response to a question, whether verbal or otherwise symbolic. However, this assumption need not be made. In open performance based tasks, the goal is the completion of a set of appropriate actions: create a dance, write a composition, choose an appropriate response, create a computer network. In the ECD framework, the language of presentation and work product is much broader than needed for questions alone. It is the role of the assessor to look at the work product and determine the relevant features in the action or work product and the value of those features as evidence for different inferences. The core idea is that when an action or work product is

needed for inference, asking a question is only one form of the task that may be available. Naturalistic assessment that occurs on the football field, in business presentations, or in assessing a paper for publication rarely takes the form of a question. Rather, a complex work product or performance is held up against rules (formal or informal, explicit or implicit) after a specific set of tools has been provided and a specific goal has been articulated. A question is one possible format.

Items Have Answers

The classic fixed response format is optimized for scoring efficiency (DiCerbo & Behrens, 2012). This is accomplished by structuring the work product to be a fixed choice that can be automatically scored by a simple comparison function between response and target answer. DiCerbo and Behrens (2012) argued that this has led to a simplification of the chain of reasoning back to the presentation process to motivate simple tasks and down the inferential chain of reasoning to the reification of statistical models that emphasize independence of observation.

However, this again is not necessary. As computer and other technologies advance, automated scoring can be applied to a broader range of actions on the work products. This can shift the focus from binary identification of a single preferred response to complex feature extraction aimed at creating observations from free-form work products, including writing essays (Dikli, 2006), configuring computers (Rupp et al., 2012), and diagnosing patients (Margolis & Clauser, 2006).

This opens the possibility not only that the work product can become more complex, but also that the types and amount of features sought can be expanded. The answer is in some ways a side effect of requiring one feature to look for; however, current computational advances release that requirement.

Items Measure Correctness

A common side effect of an item-centric view of assessment is that the assessment may be conceptualized and designed in terms of the matching algorithm of scoring as the primary conceptual lever in the assessment process. Two dangers may occur from this. The first danger is that test is conceptualized in terms of overall goodness of response based on average correctness. A common pattern for assessment design is (a) identify a domain, (b) sample ideas or activities from the domain, (c) make questions about those ideas or activities, and (d) score them as correct or incorrect. The difficulty is that this pattern can be undertaken with very little specification of the domain or discussion of the precise type of evidence or inference desired. The correctness paradigm can drive the construction with very little acknowledgment of the relationship between the role of individual items and the overall inference being sought. It begs the question: if the item is measuring correctness, I need to know correctness of what.

A second concern with the correctness paradigm is that it fails to account for the many situations in which there is interest in assessing specific attributes of an individual and not only overall goodness. There may be interest in identifying specific strategies used, the presence of a specific belief or action, or place someone in a cluster of similar individuals not because of correctness, but because of work features that are relevant to diagnosis or instruction. This is a generalized feature-centric view of response scoring. This is an important concept as work products become more ubiquitous and available for diagnostic purposes. Accordingly, we promote that the relevant KSAs of assessment are not knowledge, skills, and abilities, but rather knowledge, skills, and attributes (Behrens et al., 2012).

Items Measure One Thing

When someone speaks to you about mathematics, you infer a number of different competencies: the person can speak, knows the language to some level, knows mathematics to some level, and knows the social rules to engage you in the discussion. If he or she is talking to you on the phone, you could infer competencies related to technology use. If talking to you in person, you could infer aspects of social competence based on the person’s appearance.

This does not seem so complicated in the natural world. But one hears repeatedly about the importance of items being designed to measure one thing. In fact, this is quite impossible, as all the items that ask questions with verbal prompts are in fact measuring a number of aspects of proficiency, whether intended or unintended. In some cases, the goal of the admonishment is to ensure clarity in design to ensure construct-irrelevant activities are causing additional construct-irrelevant variance, and measuring one thing is an approach to dealing with that. However, given the integrated nature of human performance, a better view may be to be clear on what complex of activities one wants to see and instrument the tasks to collect and tease apart these strands with repeated data collection. This would require that tasks be analyzed to understand the dimensionality of the performances in terms of critical interpretive or diagnostic dimensions, as well as overall performance dimensions associated with statistical commonality.

As tasks become increasingly complex, the importance of using the increasingly complex behavior to extract increasingly rich data and inference is important. Since the introduction of the Question and Test Interoperability 1.0 assessment delivery standard, the notion of multiple observables emanating from a single task has been well established. In this scheme, a task may generate multiple pieces of data that may inform multiple independent or correlated dimensions of performance.

Figure 1 illustrates how a simple task can give one observation of information to update one proficiency estimate (a), or the information can be used to update two proficiency estimates (b). Panel (c) suggests the idea that two pieces of information can be generated by a task and provide updating information for two proficiency estimates. Panel (d) illustrates the idea that multiple observables can load differentially on multiple proficiencies. This is a model Behrens, Collison, and DeMark (2006) called the multi-observable/multistudent model variable approach.

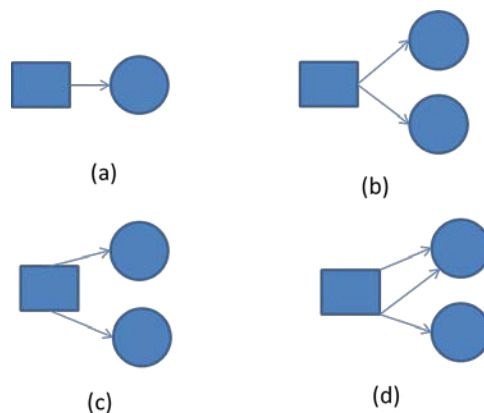


Figure 1. Multi-observable/multistudent model variable approach.

Continued advances in psychometrics build on longstanding advances in item response theory (Hambleton & Swaminathan, 2010) to model these variations using methods including diagnostic classification models (Rupp, Templin, & Henson, 2010) and Bayesian networks, or BNs (e.g., Levy & Mislevy, 2004; Levy, Mislevy, & Behrens, 2011).

ECD provides a language that allows for the description of a broad range of open natural tasks, while also describing traditional activity in classrooms and large-scale summative tests (Mislevy, Behrens, Dicerbo, & Levy, in press). We contrast this view with common assumptions of the item paradigm to remind the reader that new technologies aligned with new concepts allow for the possibility of complex assessment activities combined with complex scoring to produce rich data, allowing for meaningful inferences. To help with conceptual transition, we recommend avoiding the term *item* in favor of the more flexible and broad conceptualization of *activity*, following DiCerbo and Behrens (2012).

The discussion above focuses on the item level of analysis in assessment delivery. Following the general tenets of ECD, we believe it is important to focus on the broader conceptual units of assessment, such as the exam or the assessment ecosystem. The reader is referred to Frezzo et al. (2012) for a further discussion of these ideas in the complex simulation context.

Psychosocial Aspects of Simulation in Assessment

In all discussion of simulation, it is important to keep in mind the representational and social aspects of simulation and user interaction with simulation. By representational, we mean the symbolic and interpretive nature of a simulation as an external knowledge representation (EKR; Mislevy et al., 2007). Mislevy et al. defined EKRs as follows:

An external knowledge representation (EKR), or inscription (Lehrer & Schauble, 2002), is a physical or conceptual structure that depicts entities and relationships in some domain, in a way that can be shared among different individuals or by the same individual at different points in time. EKRs are human inventions that overcome obstacles to human information processing with respect to limited working memory, faulty long-term memory over time and in volume, coordinating the actions of many individuals, and idiosyncratic ways of thinking about some phenomenon of common interest. (Mislevy et al., 2007, p. 2)

Accordingly, EKRs are human-made and agreed upon communication and memory facilitation devices that support activity. As a special class of such devices, simulations are representations that act like or put on the appearance of other perceptual inputs. A central goal of using simulations is to create knowledge representations that act like some aspect of the world of perception and thereby induce the user to act as if the simulation could be treated as real in some way. This requires users to take on a certain epistemic position in which they attend to certain relevant aspects of their environment while ignoring other aspects, which allows them to act as if the simulation was real, and they are thereby justified to act as if they were scientists (Ketelhut, Nelson, Clarke, & Dede, 2010), or scholars, or explorers (Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007) in a simulated world.

There are two aspects of disbelief to keep in mind when building environments for assessment. First, the social situation of the exam is artificial and requires the user to enter into the assumptions of the exam. These assumptions include the idea that the work is important, should be completed, has an appropriate response, and so forth. The role-taking the learner engages in with respect to the pretend

or act as if nature of the simulations depends in part on the learner's view of the social contract governing the exam use in general.

A second aspect of the disbelief is so ubiquitous that it is easy to forget the completely socially constructed nature of the representation. Learners might easily forget that the image they are interacting with is only a set of colors that are reflecting underlying values that are mapped onto colors in the video display. The perception is only of shape and color. The goal here is create environments in which the perceptual layer can be completely interpreted as if learners were interacting with something real. This interaction may require some learning about the social expectations of treating colors on screens as psychologically concomitant with similar perceptual experiences off the screen. For learners with relatively little digital experience, this may take explicit training. On the other hand, learners with deep experience in a fixed set of electronic representations or experiences may have difficulty transferring the rules of interface interaction or social norms regarding the interpretation of icons and thereby face difficulty because of extensive prior experience. Learners who ferry between operating systems sometimes experience this awkwardness and, as other genres of digital experience evolve, other types of transfer difficulty may emerge for them.

Dimensions of Task Design and the Role of Simulation

While in the previous section we discussed the structural components of assessments and the transformational evolutions of presentation, work product, and evidence identification, a question still remains regarding the relevant psychological layers in task design. The ECD model opens one's understanding to the broad range of possible interactions and activity, but it remains silent on specific conceptualization of tasks and, therefore, how one may consider simulations from a psychosocial or activity framework.

To fill in this gap, we suggest considering tasks in terms of four aspects of the task design relevant to the end user's experience: problem space, tool space, solution space, and response space. Each of these notions will be discussed in turn with an eye for variation in roles of simulations in expanding these different categories.

By using the term *space*, we mean something analogous to how it is used in computer science or knowledge engineering literature (e.g., Stefik, 1995). In that context, for example, search space "refers to the set of symbol structures that a program can consider as candidate solutions" (p. 148) and analogous to what we refer to as solution space. In general, by *X space*, we mean the set of structures, beliefs, representations, understandings etc. that an individual (or group) can consider as candidate Xs.

Problem Space

An activity inherently holds a goal that leads to the conceptualization of the problem being solved. The problem may be informational and procedural (find the information x) or it may be affective or motivational (entertain yourself). In a testing situation, specific goals are communicated to steer the examinee toward a set of constrained actions. In the multiple-choice question, the problem space is typically highly constrained to avoid ambiguous problem interpretation that would lead to misinterpretation of the objective of the question. However, this is not universally unavoidable; therefore, psychometric review of item characteristic functions is needed to make sure the most proficient students do not bring novel interpretations of the problem space by overinterpreting the question. Of course, overinterpretation itself represents the judgment from a certain epistemic position.

In real-world tasks, the openness of the problem space may vary. In some cases, the goal and corresponding solutions are clear (add the numbers); in other cases, the problem space is left purposely ambiguous because reconstructing the problem formulation may be part of the tasks itself. Here, clarifying questions or additional research to understand linguistic cues and so forth may be the desired behavior one is seeking to elicit. On the other hand, in some situations, clarifying questions from learners meant to elicit advanced dialog about the space of interpretation may be squelched as “stupid questions” if the invigilator or teacher does not understand the problem space as appropriately having multiple interpretations.

In some cases, simulations can successfully broaden the problem space of an assessment by providing a broad environment where problem definitions can be generated. For example, instead of providing a specific goal, such as “measure the temperature at time A,” the situation may be constructed in a more open-ended way that requires either parsing or reinterpreting the activity, such as broadening the goal statement to “determine the causal factors in temperature change.”

Tool Space

Activities are accomplished with tools. In some cases, the tools are simply the intellectual capabilities of the learner; in other cases, additional affordances may be provided. Tools can be considered any affordance that potentially improves the performance of the individual over the same state without the tool.

In some cases, the basic appearance of the simulation environment is itself a tool because it may provide visual, auditory, or temporal representations, illustration, or manipulation that enhance the comprehension of otherwise complex information that would be difficult to manage in a less dynamic environment. For example, common physics simulations provide illustration and experimentation with an interactive system in ways that support movement between different scientific representations (Stieff, 2011) while allowing comprehension of complex interactions. Simulation of such tools as a chat experience, word processing, or other documentation allows the capture of relevant information in appropriate representational forms.

In other cases, simulations provide representations of standard tools that exist outside of the learning/assessment environment. For example, some simulations allow the automated (or manual) collection of data into tables or graphs in ways that mimic the tools used in scientific work. This supports elicitation of performances related to scientific practice, as well as evidence of general scientific awareness or knowledge.

The normative interpretation of tool availability is important to consider in assessment environments. If there is an assumed rule of “if the tool is there, they must want me to use it,” learners may react differently than if the rule is considered to be “sometimes they put things there to trick us.” Accordingly, different normative rules have different implications for interpreting the problem space, as tool availability may impact both problem space (it’s a spreadsheet problem) and solution space (I can use a graph or a table).

Solution Space

Another important use of simulations is to provide environments in which learners can undertake a broad range of potential avenues of activity in complex domains. This is the goal of using simulation to

broaden the solution space. By solution space, we mean the set of possible activities that can be undertaken to accomplish a goal. This is the key to the open-ended activity of unstructured daily life.

Tools offer a specific role in broadening or constraining the set of possible solutions that learners may traverse to come to their response. Forcing tool use generally constrains the space, while offering it freely broadens it. There are other methods to control the size of the solution space, including asking for a specific form of a solution or making the problem space highly constrained so only a fixed set of solutions is plausible. This is a fundamental logic of multiple-choice: limit the solutions to a fixed number and match that to a fixed set of responses.

In open-ended and performance tasks, the solution space is purposely expanded to support a wide range of activity. Complex simulators that mimic the natural world allow for a wide range of troubleshooting, problem solving, and creative activities. Activity over time allows moving about the solution space, applying and evaluating different solutions and deciding from among different solutions or expressions

Response Space

Response space is concerned with the range of possible interactions relevant to aspects of the interaction that will be evaluated. The notion of the actions relevant to evaluation is important so as to distinguish response as interaction in some aspect of the simulation versus response as creation of work products that will be evaluated. For example, some tasks exist in which simulation is provided to allow exploration of real-world problems and situations (problem space, solution space, tool space), but the indication of response occurs outside of the simulation in a multiple-choice or other format response.

For example, one item format used in the Cisco Certification program (Cisco Systems, n.d.) is called a *simlet* and is illustrated in Figure 2. A simlet consists of a network simulation that allows a broad solution and tool space (a full range of possible commands and actions in a small network) but uses a traditional response and problem space. After learners obtain information from the simulation, they bring this information to bear on the multiple-choice question.

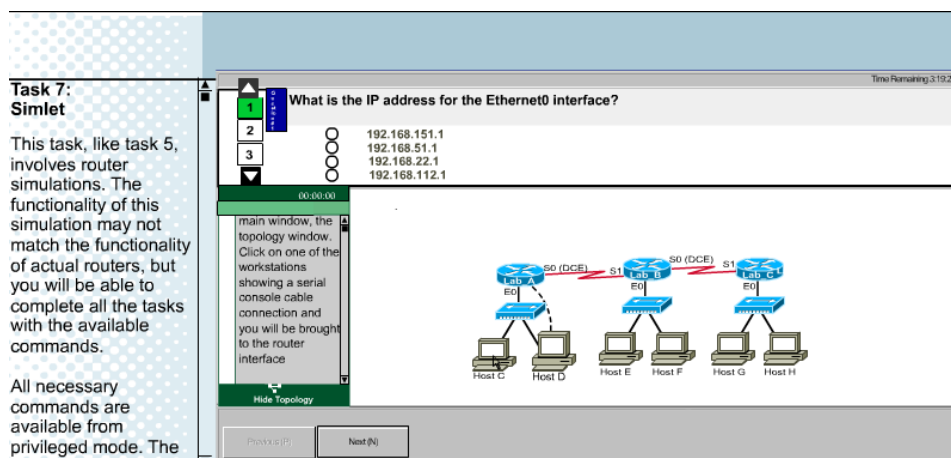


Figure 2. An example of a Cisco certification exam simlet.

In other settings, the simulation may be relatively limited in the effect of expanding the problem or solution space, but allow a broadening of the response space to allow for alignment with authentic professional activity. For example, in the simulation illustrated in Figure 3, students observe the time required for tablets to dissolve at different rates as a function of simulated water temperature. The simulation has relatively small impact on the problem or solution space, but aligns the tools space closely with the response space in the form of a graph construction.

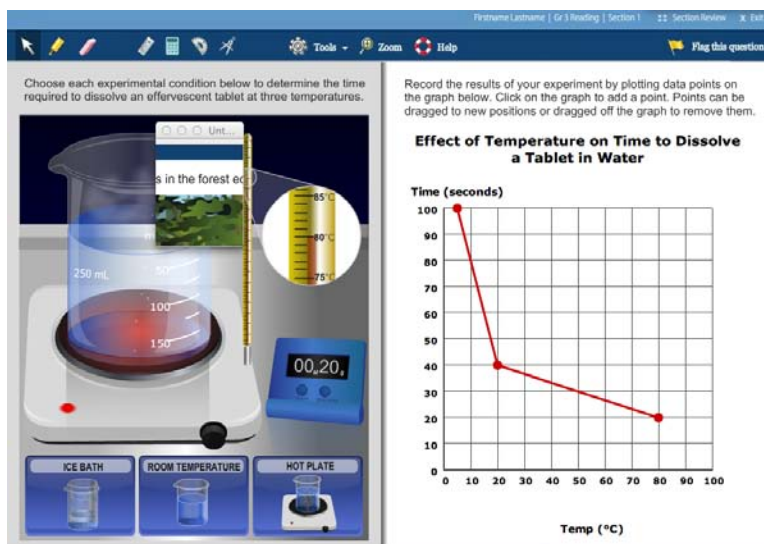


Figure 3. Example of a simulation that expands the tool space to allow interactive graphing.

Simulations are used for expanding or constraining aspects of presentation beyond methods otherwise available and to provide an environment in which individuals can act as if they were in other environments, or taking on other identities or epistemic positions. The appropriateness of a set of simulation features depends on the goals of the assessment designers and the types of affordances they expect to bring to the assessment activity, with possible variety in all four spaces (see Table 2).

Table 2. Descriptions of Four Spaces to be Manipulated in Simulations

| | Description | Narrow space | Broad space |
|----------------|-----------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------|
| Problem space | The goal or problem to be solved | What is the temperature at times A and B? | Determine the causal factors in temperature change |
| Tool space | The mechanisms that can be employed to help solve the problem | A thermometer | An array of chemistry lab equipment |
| Solution space | Possible activities that can be undertaken to accomplish a goal | Temperature readings change in 10-degree increments | Conduct experiments to determine why people in the town are getting sick |
| Response space | Activities that will result in the evaluated product | Choose one of four options | Write a letter to the town council explaining why people are getting sick |

Potential Pitfalls

When designing simulations, there are decisions to be made about what to include, the level of fidelity of the simulation, and the cognitive processes elicited. In making these decisions, there are potential pitfalls to avoid, including: oversimulating, undersimulating, uniquely simulating, passively simulating, and simulating blindly. We discuss each of these pitfalls in turn. As stated above, the interpretation of any instantiation as successful or unsuccessful depends on the precise goals of the task developer and the alignment with the assessment practice. An activity may be inappropriate for one inferential context while being appropriate for another. No activity is universally appropriate or inappropriate.

Oversimulating

When designing simulation interfaces and experiences, it is tempting to try to model every aspect of an environment. However, when the environment contains a mass of details and surface features that are not germane to the problem at hand, cognitive load may be unnecessarily increased for users without gaining any inferential value. Actions and decisions related to these details introduce construct-irrelevant variance.

Roschelle (1997) argued that high fidelity simulations may not be successful because students do not know how to make sense of them. Novice student conceptualizations may be so far from expert models that they are not able to identify the important elements of the simulation environment on which to attend. ThinkerTools (White, 1993), a Newtonian physics microworld game, embodies an approach to present simplified versions of phenomena. White designed ThinkerTools based on research about how students build knowledge rather than expert models, and ThinkerTools uses very simple user interfaces to allow students to experiment with physics concepts.

The purpose and scope of the simulation should be clear to the designer. If the goal of the simulation is to expand the problem and solution space, the designer should carefully consider whether adding increased detail or fidelity increases the problem space in construct-relevant ways or simply increases memory load or construct-irrelevant variance associated with new experience. Depending on the design goal, simulations should allow users to manipulate specific features of the environment to observe results (in a learning environment) or demonstrate understanding (in an assessment environment). If users are required to manipulate too much, it becomes difficult to model the results of their actions and for students to understand cause and effect relationships. Many business simulations, for example, require players to make large numbers of simultaneous manipulations and decisions about the business they are running, obscuring which decisions impacted their outcomes (Teach & Murrf, 2008). Micromatic, for example, has a 77-page manual (Scott, Kaliski, & Anderson, 2011) and requires upwards of 60 decisions per round, with general feedback on performance of the business at the end of the round. Figure 4 is a screen shot of just the marketing decisions required, including advertising costs by region; salesperson hires, fires, and salary structure; and location of sales teams. There are also operating, finance, and plant decisions to be made. While this is certainly a rich environment, it becomes difficult to isolate any particular skill or decision, as they are all intertwined. In addition, the potential for mental fatigue increases, and it does not reflect the real world in which these decisions would be made across teams and over time. While modeling each of these elements of the real-world environment may

seem inviting, care must be taken that it does not interfere with the ability to make inferences about constructs of interest.



Figure 4. Micromatic marketing decisions screen.

Undersimulating

Undersimulating concerns the use of simulation in a manner that does not significantly change one of the relevant presentational aspects (problem, solution, tool, or response space). If important aspects of an environment are not modeled, students will not be able to act in the environment as if they are in the real world and, consequently, will not be able to demonstrate their understanding. For example, the demonstration of color vision shown in Figure 5 allows users to manipulate the amounts of green, red, and blue reaching the person and to see the resulting color perception. Users can change the level of each color and observe the change to the man's perception of color. However, it is not clear what the man sees as purple (all the lights? something he imagines? the room?) or how vision works beyond the lesson that combining colors leads to the perception of other colors. No part of vision is modeled, so the mechanism by which perception of color occurs is not available. It would be difficult to make an observation from this simulation that would tell us about student understanding of color perception. (It should be noted that while some levels of simulation may be appropriate at one developmental level, they may be undersimulations for another level.) In the presentation space language introduced above, one may say that the solution space is broadened to allow exploration, but its relationship to the problem space is not clear.

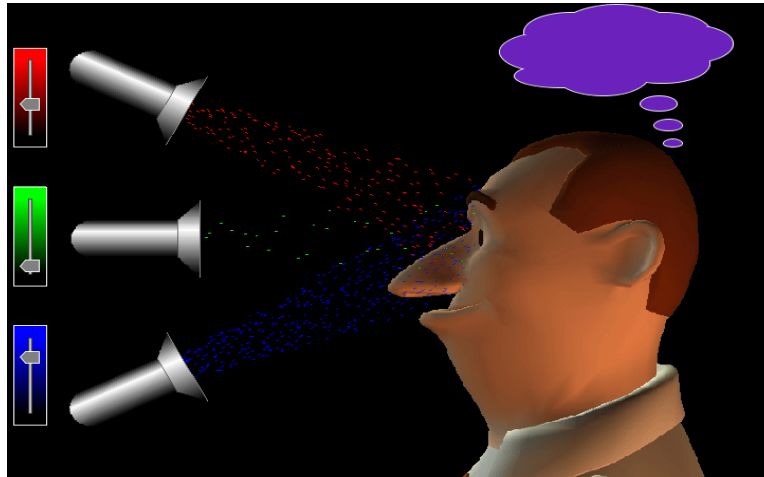


Figure 5. Color vision simulation.

Uniquely Simulating

When creating an environment or interaction, some authors create features that are unique to the simulation environment in order to simulate a certain event or phenomenon. This may be a way to interact with the interface, a particular representation of an event, or a unique perspective in a visualization. Unfortunately, every variation that is unique to the simulation and not natural to the environment being modeled introduces cause for concern. First, users have added working memory load as they try to manage the demands of the interface but still process the content of the simulation. Second, it places students with prior experience with the unique features at an advantage, creating bias in the measurement of content knowledge or skill. Both of these are then increasing the amount of construct-irrelevant variance and adding measurement error. Finally, unique features in a simulation increase the risk that knowledge and skills from the environment will not transfer outside of the environment.

An example of a unique feature in a simulation can be seen in a simulation tool from the Cisco Networking Academy called Packet Tracer (PT), which was worked on by the first two authors of this paper. In computer networking, there is a command called *ping* that tells the device to send a message to another device in order to determine whether the two devices can communicate. This command is simulated in PT from the command line, as it is in the real world (see Figure 6). However, students can't actually visualize what is happening during the ping issued from the command line. In addition, the scoring mechanisms were not able to determine whether the ping was successful from the command line (i.e., whether the student had correctly gotten the devices to communicate).

To address these issues, an alternative, called a protocol data unit (PDU), was developed that allowed for a graphical representation of the ping (see Figure 7) and also allowed the scoring mechanisms to determine whether communication was successfully established. However, this form of a PDU is unique to Packet Tracer. In user testing of a simulation that required students to send the PDU, students spent unacceptable amounts of time trying to figure out what to do in the interface in response to the instructions. The uniqueness of the simulation became a barrier to completing the task.

Invitational Research Symposium on Technology Enhanced Assessments

In addition, other students began to use the PDUs in place of the ping command in all situations, raising the concern that they would not be fluent with the ping command when using real devices.

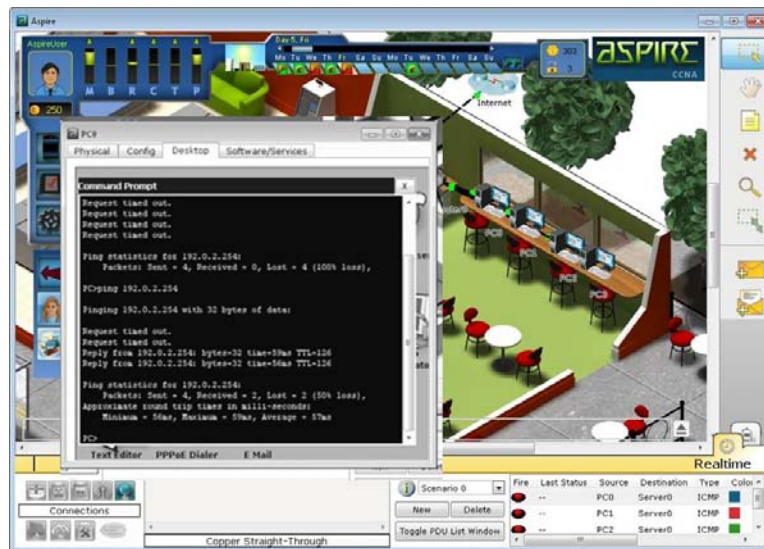


Figure 6. Command line interface.



Figure 7. Unique PDU model.

Passively Simulating

One of the benefits of simulation is that students can interact with the elements of the simulation and either test their theories of the world (exploring the solution space) or demonstrate their understanding (leveraging the response space). If a simulation does not allow for students to interact with a simulation in significant ways, it is difficult to make inferences about what they know and can do. While multiple-choice questions can be used to follow up, students are then being asked about an animated illustration rather than making full use of a simulation tool as a way to gather evidence. A review of algorithm visualization studies (Hundhausen, Douglas, & Stasko, 2002) found that experiments that manipulated learner activities around the representations were more likely to show impact on learning than experiments that manipulated the representations. In other words, what the students did in the simulation was more important than the representation itself. Further, students who just viewed the visualizations did not demonstrate learning gains.

To clarify the difference between passive and active simulations, look at the website of the Chemistry Education Research Group at Iowa State University (n.d.). The group makes a specific distinction between simulations and animations. For example, under the topic of solutions, there is an animation showing how water molecules act to dissolve a salt molecule. In this section, the learner can replay the animation and click through different views (see Figure 8) but is not involved in setting any parameters of the simulation itself. In contrast, there is a simulation of a conductivity test (see Figure 9) in which students select the solution, volume, and weight and then run a conductivity test. In this case, the student can actively manipulate the test itself.

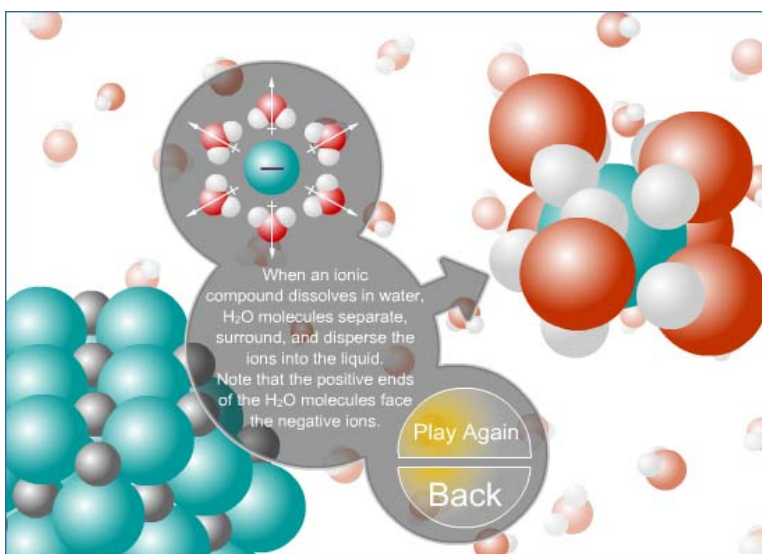


Figure 8. Animation related to chemical solutions.

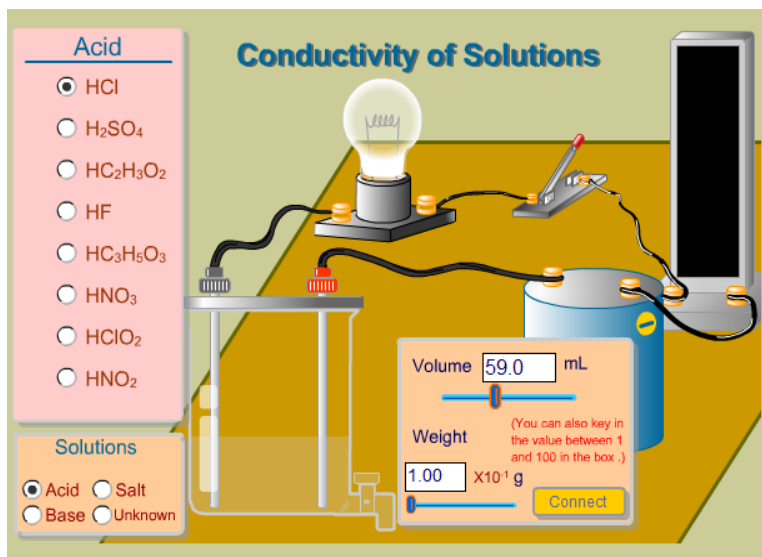


Figure 9. Simulation related to chemical solutions.

Simulating Blindly

When creating simulations for assessment, one needs to be clear about the cognitive processes that are invoked by the simulation. If simulations are to be used to make inferences about student knowledge, skills, and abilities, we must be clear what processes are used to interact with a simulation. If students interacting with a chemistry simulation are creating molecules based on known rules of combination, one makes inferences about their knowledge of such rules based on their final product. If they are able to join simulation elements and arrive at the correct answer by trial and error, one will have more difficulty making inferences from that result. This concept of the processes students use to complete tasks is at the heart of substantive validity (Messick, 1995).

However, in some cases, it is not clear how students are solving problems in a simulation environment. For example, the Napoleonic Wars OnLine (NWOL; Historical Online Learning Foundation, 2008) is a multiplayer simulation of combat during the Napoleonic Wars. At the beginning of each turn of an NWOL game, players receive reports showing where their units and ships are, what they can see, and what happened on the previous turn. Players then submit orders for their units via a web form, in consultation with their fellow players. Game boards (see Figure 10) are used to track players' moves. Although the game is not designed specifically for classroom use, it is recommended in lists of history simulation games for educators (e.g., Social Studies Central, n.d.). However, there is no evidence for the processes players use to decide what orders to give. Are they based on historical knowledge, military strategy, game theory, or best guess? Without knowing, there is no way to make inferences from a student's performance in the simulation to what they know and can do. Again, it should be emphasized that the simulation is not initially intended to provide these inferences, but if the suggestion is taken to use it in the classroom, these questions must be explored.

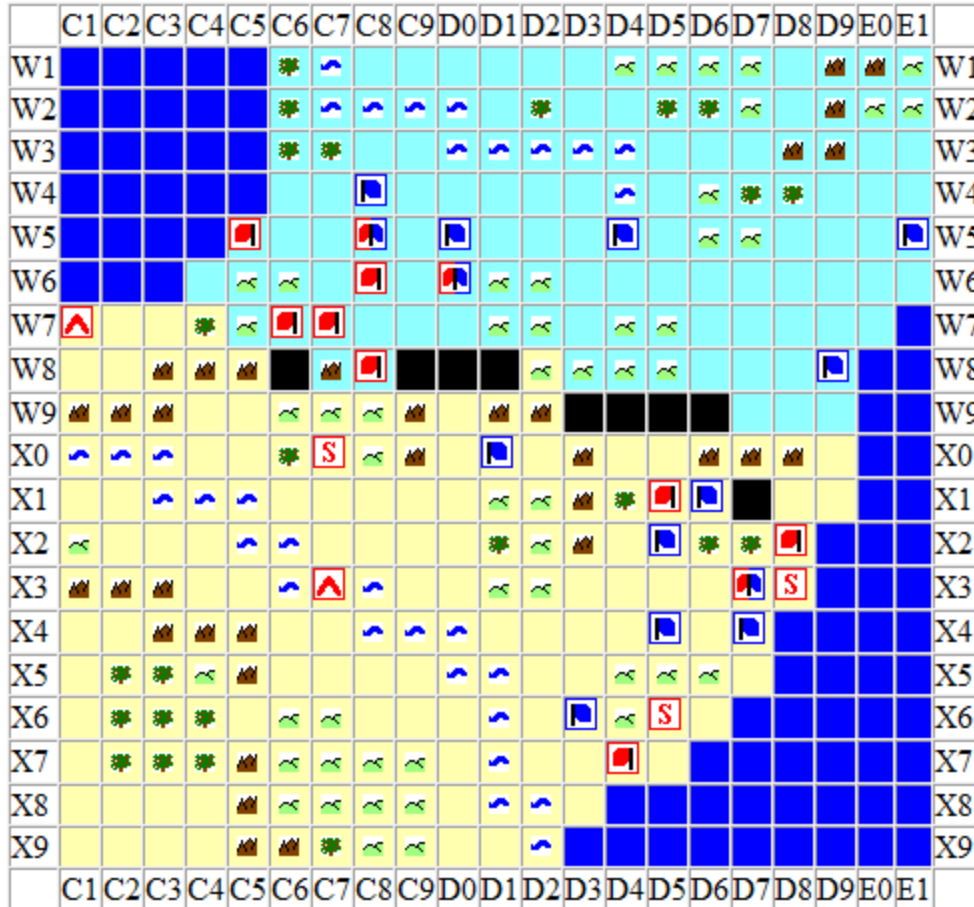


Figure 10. Game board from Napoleonic Wars OnLine.

Summary and Conclusions

Traditional assessment methods have evolved a standard operating procedure and language that obscures some of the possible richness available for assessment made possible by advances in interactive computer technologies and scoring, as well as improved understanding of the psychosocial issues related to assessment development. The ECD terminology opens the door to a broader set of conversations and conceptualizations while the presentation space framework presented here attempts to open the conversation to alignment of purpose and cognitive value of simulations.

Following the general evidentiary logic of assessment, we discussed a number of ways in which simulations offer the possibility of deceiving learners into acting in ways consistent with educators' hopes and expectations by providing engaging environments that invite the learner to deceive themselves into acting as if the simulation were real in a larger context. At the same time, we caution assessment designers to consider which dimensions of realism or presentation change they are seeking

to expand and to avoid deceiving themselves into believing the value of their interactions simply because more interaction is provided.

Author Notes

John T. Behrens is Vice President, Center for Digital Experience and Analytics at Pearson and Adjunct Assistant Research Professor in the Department of Psychology at the University of Notre Dame, Notre Dame, IN, USA. John.Behrens@pearson.com

Kristen DiCerbo is Senior Researcher Scientist, Center for Digital Experience and Analytics at Pearson. Kristen.DiCerbo@pearson.com

Steve Ferrara is Vice President, Center for Performance Assessment at Pearson.
Steve.Ferrara@pearson.com

References

- Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *Journal of Technology, Learning, and Assessment*, 1(5). Retrieved from <http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1671>
- Barab, S. A., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: A framework for socio-scientific inquiry. *Journal of Science Education and Technology*, 16, 59–82.
- Behrens, J. T., Collison, T. A., & DeMark, S. F. (2006). The Seven Cs of comprehensive assessment: Lessons learned from 40 million classroom exams in the Cisco Networking Academy Program. In S. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies in higher education, K-12 and corporate*. (pp. 229–245). Hershey, PA: Information Science Publishing.
- Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfek, & H. F. O’Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). New York, NY: Erlbaum.
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.). *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–54). Charlotte, NC: Information Age Publishing.
- Chemistry Education Research Group at Iowa State University. (n.d.). *Chemistry experiment simulations and conceptual computer animations*. Retrieved from <http://group.chem.iastate.edu/Greenbowe/sections/projectfolder/simDownload/index4.html>
- Cisco Systems. (n.d.). *Certification Exam Tutorial*. Retrieved from http://www.cisco.com/web/learning/le3/learning_certification_exam_tutorial.html
- DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1). Retrieved from <http://ejournals.bc.edu/ojs/index.php/jtla/index>
- Frezzo, D. C., Behrens, J. T., DiCerbo, K. E., & Chen, M. (2012). An extensible micro-world for assessment of activities in the data networking professions. Paper presented at the National Council on Measurement in Education Annual Meeting, Vancouver, BC.
- Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology*, 19, 105–114.
- Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.
- Historical Online Learning Foundation. (2008). *Napoleonic Wars OnLine*. Retrieved from <http://holf.org/NWOL/index.htm>
- Hundhausen, C. D., Douglas, S. A., & Stasko, J. T. (2002). A meta-study of algorithm visualization effectiveness. *Journal of Visual Language and Computing*, 13, 259–290.

- Ketelhut, D. J., Nelson, B. C., Clarke, J. E., & Dede, C. (2010). A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology, 41*, 56–68.
- Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a simulation-based assessment. *International Journal of Testing, 4*, 333–369.
- Levy, R., Mislevy, R. J., & Behrens, J. T. (2011). Markov chain Monte Carlo in educational research. In A. Gelman, G. Jones, X. L. Meng, & S. Brooks (Eds.), *Handbook of Markov chain Monte Carlo: Methods and applications*. London, England: Chapman & Hall/CRC Press.
- Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. W. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–168). Mahwah, NJ: Lawrence Erlbaum.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from a person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
- Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., Robinson, D. H., Rutstein, D. W., Shute, V. J., Stanley, K., & Winters, F. I. (2007). On the roles of external knowledge representations in assessment design (CSE Report 722). Los Angeles, CA: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved from <http://www.cse.ucla.edu/products/reports/r722.pdf>
- Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (in press). Data mining versus psychometrics in educational assessment: An evidence centered design approach. *Journal of Educational Data Mining*.
- Roschelle, J. (1997). Designing for cognitive communication: Epistemic fidelity or mediating collaborative inquiry? In D. L. Day & D. K. Kovacs (Eds.), *Computers, communication, and mental models*. Bristol, PA: Taylor & Francis.
- Rupp, A. A., Levy, R., DiCerbo, K. E., Benson, M., Sweet, S., Crawford, A. V., Fay, D., Kunze, K. L., Caliço, T., & Behrens, J. T. (2012). The interplay of theory and data: Evidence identification and aggregation for product and process data within a digital learning environment. Manuscript submitted for publication.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
- Scott, T. W., Kaliski, J. A., & Anderson, P. H. (2011). *Micromatic student manual version 4.0*. Retrieved from <http://oaktreesim.com/micromatic/manuals/MicromaticStudentManual.pdf>
- Shaffer, D. W., & Gee, J. P. (2012). The right kind of GATE: Computer games and the future of assessment. In M. Mayrath, D. Robinson, & J. Clarke-Midura (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 211–228). Charlotte, NC: Information Age Publications.
- 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., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education, 18*(4), 289–316.
- Social Studies Central. (n.d.). *Online interactive simulations*. Retrieved from <http://socialstudiescentral.com/?q=content/online-interactive-simulations>
- Stefik, M. (1995). *Introduction to knowledge systems*. Waltham, MA: Morgan Kaufmann.
- Stieff, M. (2011). Improving representational competence using multi-representational learning environments. *Journal of Research in Science Teaching, 48*, 1137–1158.
- Teach, R., & Murrf, E. (2008). Are the business simulations we play too complex? *Developments in Business Simulation and Experiential Learning, 35*, 205–211.
- Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- White, B. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10*, 1–100.
- Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., & Behrens, J. T. (2004). Design rationale for a complex performance assessment. *The International Journal of Testing, 4*, 303–332.



Invitational Research Symposium on
Technology Enhanced Assessments

The Center for K–12 Assessment & Performance Management at ETS creates timely events, where conversations regarding new assessment challenges can take place, and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state, and local decision makers.

Copyright 2012 by Pearson.

ETS is a registered trademark of Educational Testing Service (ETS).