K-12 Center at ETS

Invitational Research Symposium on
Technology Enhanced Assessments

Psychometric Advances, Opportunities, and Challenges for Simulation-Based Assessment

Roy Levy

May 7–8, 2012

K-12 Center at ETS
Invitational Research Symposium on
Technology Enhanced Assessments

Psychometric Advances, Opportunities, and Challenges for Simulation-Based Assessment

Roy Levy

Arizona State University

Executive Summary

This report characterizes the advances, opportunities, and challenges for psychometrics of simulation-based assessments through a lens that views assessment as evidentiary reasoning. Simulation-based tasks offer the prospect for student experiences that differ from traditional assessment. Such tasks may be used to support evidentiary arguments commonly used in assessment. Importantly, they also support evidentiary arguments that differ considerable from those typical in assessment. These novel assessment arguments are richer or more nuanced than those commonly used in terms of the targeted inferences about students, the evidence that facilitates those inferences, and the tasks that allow for the collection of such evidence. In the adopted framework that views assessment as an evidentiary argument, a psychometric or measurement model serves as the machinery to facilitate inference from observing student actions to beliefs about their proficiencies, skills, knowledge, misconceptions, strategies, and so on. We review recent advances in statistical modeling that offer a variety of psychometric models to these effects. It is argued that innovative measurement modeling frameworks, though not as well developed as those that dominate current operational assessment, are well poised to handle the complexities of the rich assessment arguments supported by innovative simulation-based assessment. Related aspects, strategies, and pitfalls in other parts of the assessment development process that must cohere with the psychometric model are discussed, including those surrounding assessment design, data analysis, and assessment and model revision. It is further argued that many of the key challenges facing psychometrics of simulation-based assessment are best resolved by a principled approach to assessment design in concert with data analysis. Strategies for solutions to some of the more imminent challenges to psychometrics for simulations are discussed, and potential short- and long-term future developments are suggested.

Introduction

Advances in computing allow for new opportunities for assessment, with influences on activities and tasks, delivery systems, data collection, and data analysis. Digital representations and computer-based delivery capacities have opened up possibilities for simulation-based assessments that were previously impractical in large-scale settings. Accordingly, simulation-based assessments and related environments, such as game-based assessments and intelligent tutoring systems, that employ simulation-based assessments are receiving an increasing amount of attention. However, in some cases, the rush to adopt the *physical* machinery has outstripped the adoption of the *assessment* machinery required to sensibly use the physical technology for assessment ends, often producing disastrous (and expensive) results.

This report focuses on one aspect of assessment machinery necessary to best support the use of simulation-based assessments. Specifically, we characterize the advances, opportunities, and challenges for psychometrics of simulation-based assessments. We do so first by couching psychometric modeling and analysis in the broader context of assessment inference and argumentation offered by evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003), whereby a psychometric or measurement model serves as the machinery to facilitate inference from observing student actions to beliefs about their proficiencies, skills, knowledge, misconceptions, strategies, and so on. Two following sections then clarify the scope of what will be considered as psychometrics and review psychometrics in traditional assessment. We introduce some representative examples of simulation-based assessments and then turn to a deeper discussion of the core ingredients of psychometric modeling and the advances, opportunities, and challenges afforded by simulation-based environments. Included in this discussion will be implications for related aspects of assessment development and use. We then list some additional challenges to psychometric modeling for simulation-based assessments and offer a prospective look at the potential future use of psychometrics for simulations. A brief summary highlighting the key themes concludes the paper.

A Synopsis of Evidence-Centered Design

In this section, we review ECD to locate aspects of psychometric and measurement modeling in the larger assessment process and explicate its connections with the other components.

ECD is a framework that lays out the fundamental entities, their connections, and actions that take place in assessment (Mislevy et al., 2003; see Behrens, Mislevy, DiCerbo, & Levy, 2011 and Mislevy, 2012 for presentations with a focus on simulation-based and related environments). It provides terminology and sets of representations (Mislevy et al., 2010) for use in designing, deploying, and reasoning from assessments. It is *descriptive* in the sense that all forms of assessment (e.g., simulation-based performance tasks, multiple choice tests, teacher–student conversations) may be framed in terms of ECD, regardless of their varying surface features. ECD is also *prescriptive* in the sense that it provides a set of guiding principles for the design of assessment, the aim of which is to articulate how the assessment can be used to make warranted inferences, that is, how assessment is an instance of evidentiary reasoning. Importantly, it is *not* prescriptive in the sense of dictating what particular forms, models, and representations to use. ECD (a) helps us understand the argumentation behind the use of measurement models like those found in item response theory (IRT; Hambleton & Swaminathan, 1985); (b) helps us through the assessment development process, which might lead to the use of such IRT

models; but (c) does not *require* the use of such models. This is crucial, as the focus of the majority of this report is on newly emerging psychometric models, and the larger point is that ECD is a common foundation on which we can build psychometric models and other features of simulation-based assessments.

A quotation from Messick (1994) is useful in understanding the idea of an assessment argument and the perspective of ECD:

> We would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16)

In the following, we briefly review the layers of ECD, depicted in Figure 1, before turning to the central focus of this report.

Domain Analysis and Domain Modeling

In domain analysis, we define the content of the domain(s) to be assessed, namely, the subject matter, the way people use it, the situations they use it in, and the way they represent it. Resources here include subject matter expertise, task analyses, and surveys of artifacts (e.g., textbooks). In domain modeling, the information culled in domain analysis is organized in terms of relationships between entities, including observable behaviors we might see people do that constitute evidence of proficiency and in what situations those actions could occur or be evoked. It is during this stage that we essentially lay out the assessment argument that will be instantiated when tasks are built, the assessment is delivered, student performances are scored, and inferences are made. This often involves articulating what claims we would like to be able to make, Toulmin diagrams for assessment arguments, and design patterns for tasks.

Conceptual Assessment Framework

In the conceptual assessment framework (CAF), we take the results from domain modeling and, in light of the purposes and constraints, we devise the blueprint for the assessment. Our focus will be on the three central models that address the questions articulated by Messick (1994), as quoted earlier. What are we measuring, and what are the desired inferences? What behaviors might we observe that would constitute evidence in support of measuring those constructs or making those inferences? What situations should students be in to yield those observations? We address these questions via a few key models that become the blueprint for jointly developing the tasks and psychometric models. Three of the central models in the CAF are depicted in Figure 2. We will treat each in turn here briefly; the discussion of the psychometrics for simulation-based assessments will trade heavily on viewing what goes on in each of these models.

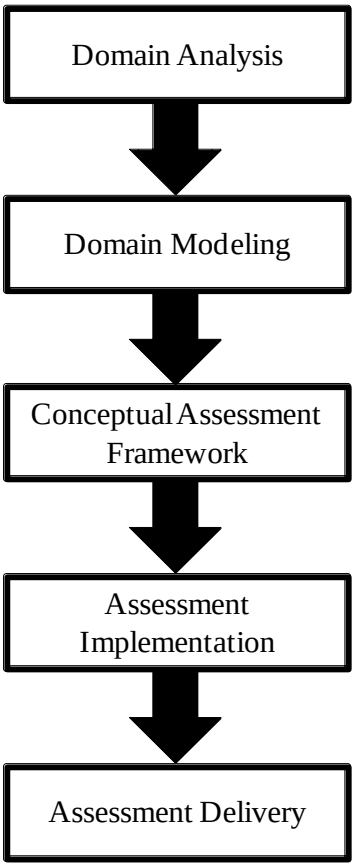

Figure 1. Layers of ECD.

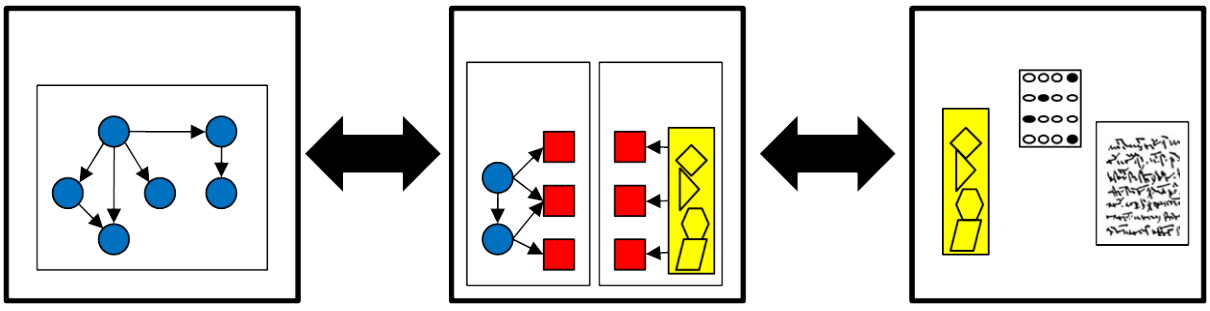

Figure 2. Three central models of the conceptual assessment framework.

Student model. The student model addresses the first of Messick's questions and is where we articulate what we are measuring and what the desired inferences are. The student model lays out the relevant configuration of the aspects of proficiency to be assessed and therefore represents choices regarding what aspects of the domain identified in the domain analysis and domain modeling will be included in the assessment. We build the student model by specifying what have variously been termed student model variables (SMVs), proficiency model variables, or competency model variables. As discussed subsequently, modern psychometric models are characterized by their use of *latent* variables as SMVs, which reflects that ultimately, what is of inferential interest about students cannot be measured directly; that is, in assessment, we seek to reason from what students say, do, or produce in certain environments to their proficiency more broadly viewed.

Task models. Task models answer the last of Messick's questions by specifying the situations in which relevant evidence can be collected to inform the values of the variables in the student model and thereby yield inferences about students. Task models specify two main aspects: What are the features of the tasks, activities, or situations presented to the student? And what are the work products—the things students say, create, or do in these situations—that will be collected?

Evidence models. As depicted in Figure 2, an evidence model connects the work products collected as specified in the task model to the variables in the student model. This is achieved via the two ingredients of *evidence identification rules* and the *psychometric* or *measurement model*. Evidence identification rules declare how the work products will be evaluated. The results of applying evidence identification rules to work products are values for observable variables (OVs). A psychometric or measurement model then specifies the relationships between the OVs and the SMVs. Modern psychometric modeling is characterized as specifying these relationships by modeling the OVs as stochastically dependent on latent SMVs (Almond, Williamson, Mislevy, & Yan, in press; Bollen, 1989; Hambleton & Swaminathan, 1985; McDonald, 1999; Rupp, Templin, & Henson, 2010), examples of which are discussed in subsequent sections. This represents a departure from classical test theory (CTT) approaches that do not formally include latent variables (Lord & Novick, 1968), though it is noteworthy that CTT may be advantageously reconceived in latent variable modeling frameworks (Bollen, 1989; McDonald, 1999). In the current work, we focus on latent variable measurement models.

Assessment Implementation and Assessment Delivery

In assessment implementation, we manufacture the assessment, including authoring the tasks; building automated extraction; parsing and scoring processes to move from work products to OVs; and forming the statistical models. Assessment delivery can be described as a four-process architecture (Almond, Steinberg, & Mislevy, 2002), cycling through the steps of (a) task or activity selection (i.e., what should the student work on next?); (b) task or activity presentation (i.e., delivery of the chosen task); (c) evidence identification through the application of evidence identification rules, in which the student work products are evaluated to produce values for OVs; and (d) evidence accumulation, which occurs when values for the OVs are entered into the measurement model to produce estimates or updated values for the SMVs.

The Scope of Psychometrics

The term *psychometrics* has frequently been used to refer to the measurement model component of the evidence model. From this view, other aspects of assessment (task creation, evidence

identification rules, etc.) are thought to lie outside the bounds of psychometrics and fall under the purview of content experts and other members of the assessment team. Slightly broader conceptions may include evidence identification as part of psychometrics, a slightly broader view might involve more aspects of the CAF, and a still broader view might include aspects of the four-process framework of assessment activity (e.g., task selection in adaptive testing is sometimes thought of as within the purview of psychometrics) or other layers of ECD. Though there are considerable benefits to taking a broad view of psychometrics (Mislevy, Behrens, DiCerbo, & Levy, in press), in the balance of this report, the characterization of psychometrics will focus heavily on evidence models and the measurement model in particular. Importantly, the connections between the measurement models and the rest of assessment development are crucial, and therefore the discussion of psychometrics necessarily includes other aspects of assessment. The importance of considering measurement models jointly with other aspects of assessment will be a recurring theme.

The motivation of this focus is grounded in a view of the measurement model as the distillation of the assessment argument into formalizations that are the machinery of inference (Mislevy, 1994). In particular, the measurement model is the junction point where the student's behaviors (captured by OVs defined by evidence identification rules) are used to update our beliefs about the student's proficiency (captured by the SMVs). To explicate, we take the simple situation where an OV *X* can take on values 1 and 0 (e.g., as the result of applying a correct–incorrect evidence identification rule) and a SMV θ can take values of high or low (corresponding to a simple conception of proficiency). The rows in the Table 1 give an example measurement model, with a conditional probability distribution for values of *X* given the value of θ.

Table 1. Conditional Probability Table for an Observable Variable *X* Given a Student Model Variable θ

Student model variable θ	Observable variable *X*	
	1	0
High	.9	.1
Low	.2	.8

By constructing the model such that *X* is stochastically dependent on θ, we specify the model as one with a "flow" from θ to *X*. This is also reflected in the graphical representation depicted in Figure 3.

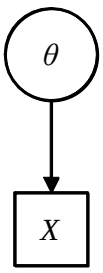

Figure 3. Graphical representation of a measurement model with one observable variable *X* and one student model variable θ.

In assessment-based inference, we seek to argue in the reverse direction of this flow, from an observed value for *X* back to θ. That is, we seek to draw inferences from observations of student behavior to their latent proficiency.

The use of probabilities in the measurement model reflects our uncertainty regarding the connection between performance and proficiency, as is almost always present in assessment (Mislevy, 1994; Mislevy & Levy, 2007). As a consequence, we have uncertainty in our inferential argument. We cannot be sure that a student who correctly answered a question has a high level of proficiency (e.g., he or she may be of low proficiency but guessed correctly), and we cannot be sure that a student who incorrectly answered the question has a low level of proficiency (e.g., he or she may be of high proficiency but slipped up). Our inference is then characterized by (a) setting up the flow from proficiency to performance—from SMV(s) to OV(s)—and then (b) reversing this direction once values for OVs are known. Returning to the example in Table 1, this is accomplished statistically as follows. For any observed value (1 or 0) for *X,* we utilize the corresponding column as a likelihood function for updating beliefs about the student proficiency represented by θ. For example, if we observe $X = 1$ (i.e., the student correctly answered the question), our beliefs regarding θ are updated by using the likelihood of 9:2 in favor of the student being at a high level of θ.

As advocated by ECD, we conceive of this process—where the model is built with a flow in a certain direction, inference proceeding by reasoning the reverse direction "back through" the model, all the while using probabilities as expressions of the strength of relationships and our uncertainty—as one of Bayesian inference. However, the argument structure still applies for variants of these themes, such as the use of frequentist perspectives on statistical inference or deterministic relationships.

Psychometrics of Traditional Assessment

Before turning to simulation-based assessment, it is worth describing traditional assessments in terms of their psychometric underpinnings as they play out in the evidence models and related aspects of the CAF. By traditional assessment, we do not mean (just) the physical objects with which students interact but also what Mislevy et al. (in press) characterized as the *standard assessment paradigm*. This may be briefly summarized as having the following elements: a student receives many short, predefined tasks (usually questions), his or her behaviors on each task (usually responses to questions) are scored independently, and those scores are aggregated to yield a summary score interpreted as a summary of the student's proficiency in the domain. Though many exceptions to this paradigm exist, this represents the dominant approach to educational assessment, particularly in large-scale and high-stakes applications.

In what follows, we adopt unidimensional IRT as an archetypical measurement model for traditional assessments, which will later serve as a foil to novel alternatives that are potentially better suited for simulation-based assessments. However, any such comparison of the relative merits and demerits of these measurement models depends on the inferential frame represented by the remaining elements of the student, evidence, and task models. Put another way, the popularity of familiar measurement models in operational assessments, particularly in large-scale assessments, reflects the purposes, choices made, and practical constraints in such assessments (DiCerbo & Behrens, 2012).

Student models in such assessments have tended to be unidimensional, where evidence accumulates to characterize students in terms of one broadly conceived construct, dimension, or

representation of proficiency, denoted here by θ. Conceptually, θ represents a single dominant dimension that drives performance on the tasks. Its interpretation rests on the scope of the tasks, which have tended to be short, discrete questions or activities. Domain modeling typically identifies a number of aspects of proficiency; design patterns and task patterns may then be used to build tasks that target these different aspects. The interpretation of θ is then derived from assessment assembly rules that guarantee inclusion of many tasks that broadly survey the domain and target the varied aspects of proficiency. This is accomplished by adhering to content specifications in fixed forms or content requirements/constraints in adaptive testing scenarios.

Work products take the form of paper-and-pencil submissions or recorded inputs in computer-based delivery systems. Evidence identification rules to produce OVs for such work products for short tasks in educational assessment has been dominated by dichotomous scoring. An archetypical example is the Scantron sheet work product with dichotomous scoring based on whether the selected (bubbled in) response to the multiple choice question is deemed correct. Partial credit scoring, often associated with rated or constructed response tasks, is similarly popular.

Measurement models have included fairly straightforward aggregations of OVs to locate students on the single SMV of the (unidimensional) student model such as via IRT models. Figure 4 depicts the measurement model graphically, where θ_i is the latent SMV for student *i* and the rectangles for X_{i1}, X_{i2}, X_{i3}, . . . , X_{iJ} are the *J* (dichotomous) OVs for student *i*, values of which are obtained by applying the evidence identification rules to the work products. In simple dichotomous scoring, there is one observable produced for each task; for each task $j = 1, \ldots, J$, observable X_{ij} is 1 if the response to task *j* from student *i* is evaluated as correct, and is 0 otherwise. When discussing a variable without reference to a particular student, we will drop the subscript *i*. The arrows from the SMV to the OVs indicate that the measurement model is constructed by specifying the distribution for each OV as conditional on the SMV. IRT models structure this dependence in terms of parameters for items. One common representation of the three-parameter logistic (3PL) model (Hambleton & Swaminathan, 1985) specifies the conditional distribution for student *i* for observable *j* (i.e., corresponding to the scored performance on task *j* by student *i*) as

$$P(X_{ij}=1\,|\,\theta_i,a_j,b_j,c_j)=c_j+(1-c_j)\frac{\exp[Da_j(\theta_i-b_j)]}{1+\exp[Da_j(\theta_i-b_j)]}$$
$$P(X_{ij}=0\,|\,\theta_i,a_j,b_j,c_j)=1-P(X_{ij}=1\,|\,\theta_i), \qquad (1)$$

where X_{ij} is the value for the OV (coded as 0 or 1); a_j, b_j, and c_j, are discrimination, location, and lower-asymptote parameters for observable (item) *j*; and *D* is a scaling coefficient usually chosen to be 1, in which case, it drops out of the expression, or 1.7, which yields results in a metric close to normal-ogive models (cf. McDonald, 1999, for other representations useful for exploiting connections to other latent variable models). Popular special cases include the 2PL model, in which $c_j = 0$, and the one-parameter logistic model, in which, additionally, all a_j are equal and, without loss of generality, can be set equal to 1 if the variance of the θ distribution is unconstrained.

This approach to measurement modeling emphasizes the use of many discrete OVs, corresponding to scores from the administration of many short tasks. Importantly, a key assumption of these measurement models is that the OVs are viewed as conditionally (locally) independent given the latent SMV θ (Hambleton & Swaminathan, 1985; McDonald, 1999). Graphically, this is represented in Figure 4 by θ_i "blocking" the path from one OV to another (Pearl, 1988). Statistically, this conditional independence assumption allows for the factorization of joint distributions of OVs. This supports a variety of activities such as simplified model calibrations and parameter estimation, test assembly to reflect target information functions, and updating beliefs about θ based on individual observations in adaptive testing. Conceptually, this conditional independence assumption means that once the evidentiary impact of a value for an OV is included by updating beliefs about θ, the OV has no further influence on the evidentiary bearing of any other OVs.

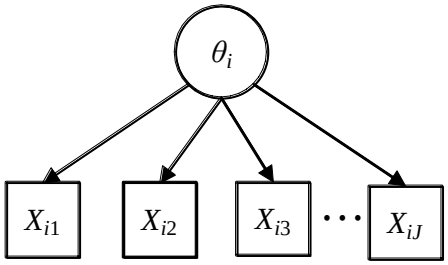

Figure 4. Graphical representation of a traditional unidimensional measurement model.

Importantly, these features of the student, task, and evidence models function as a coherent set. The key assumption of conditional (local) independence between responses in unidimensional IRT models is aligned well with the specification of a single SMV and the use of many short, discrete tasks that can be interchangeably swapped in and out of operational use.

We do not claim that all educational assessments make these choices or instantiate the inferential argument with these choices for the student, task, and evidence models. Many familiar assessments (e.g., classroom teachers questioning students; physicians engaging in medical diagnosis) depart from these choices, to say nothing of innovative assessment such as simulation-, game-, and intelligent tutoring–based assessments. Nevertheless, all these assessments can be framed in terms of the language and layers of ECD. Thus ECD allows us to see similarities and differences among various kinds of assessments that differ in their surface features (e.g., how is a game like and unlike more traditional assessments? Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008).

Simulation-Based Tasks and Assessments

A simulation-based task is one in which the student is presented with, works with, or produces a work product that contains a simulation of a real-world scenario. They are distinguished by the presence of dynamic or interactive features (e.g., viewing an animation, acting in ways that prompt a change or response from a system). They are often delivered and store work products via computer or similar digitized mechanisms, though this need not be the case (e.g., Dillon, Boulet, Hawkins, & Swanson, 2004; Vargas, 2012).

To make a number of issues concrete, we briefly describe a few environments to give a sense of the sorts of things meant by simulation-based tasks. Figure 5 contains a screenshot of Cisco's Packet Tracer system, a dynamic visualization and computation engine for simulating computer networking environments. Packet Tracer is a tool provided by the Cisco Networking Academy (Cicso, n.d.), in which information technology is taught through combinations of face-to-face instruction, online curricula, and online assessments, providing training to over one million students annually in more than 170 countries. Our current focus is on the use of Packet Tracer as a flexible environment for designing, delivering, and scoring simulation-based tasks. Tasks include activities like configuring networks to meet client needs and troubleshooting malfunctioning networks. Students engaging with Packet Tracer tasks interact with authentic images and representations of the physical devices, their interiors, ports, and so on (Figure 6), as well as an authentic command line interface for interacting with these devices. In the topological map interface (Figure 5), devices are represented as icons. A variety of toolbars allow for students to select devices and simulate the movement of packets throughout the network, which is represented as an animation. Clicking on devices brings up visuals of their physical representation (Figure 6) as well as simulations of their interfaces. See Frezzo, Behrens, and Mislevy (2010) for a more detailed discussion of the use of Packet Tracer for assessment and Cisco (2010) for a brief video demonstration.

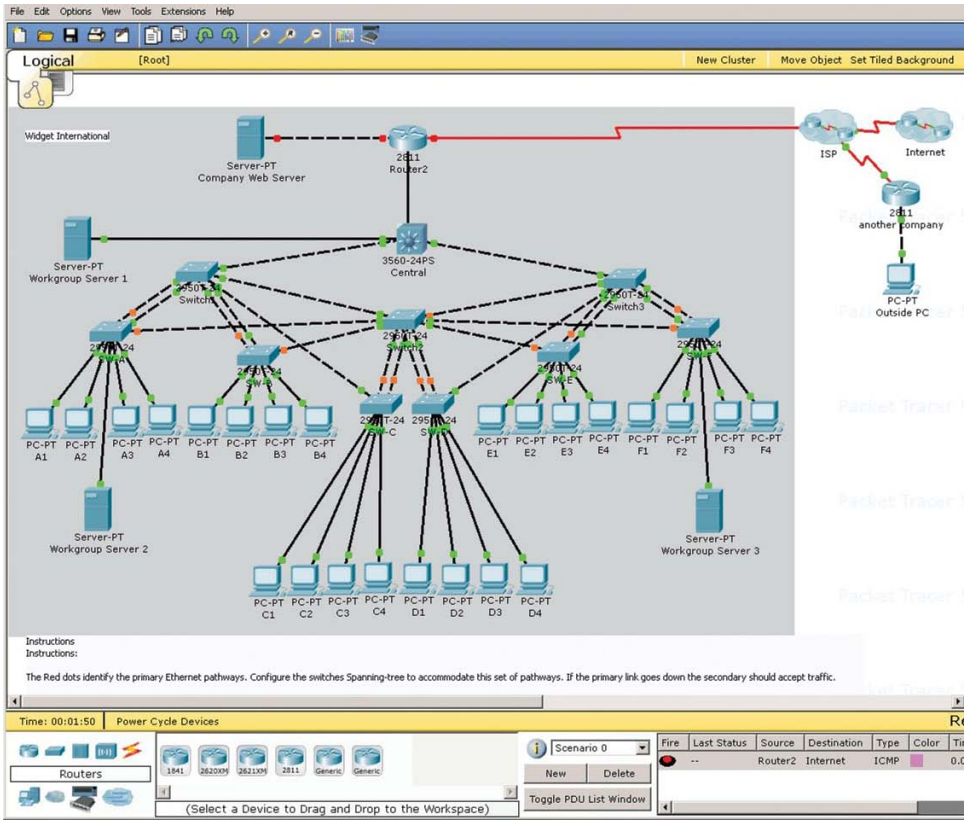

Figure 5. Screenshot of Packet Tracer, showing the logical topology and toolbars.

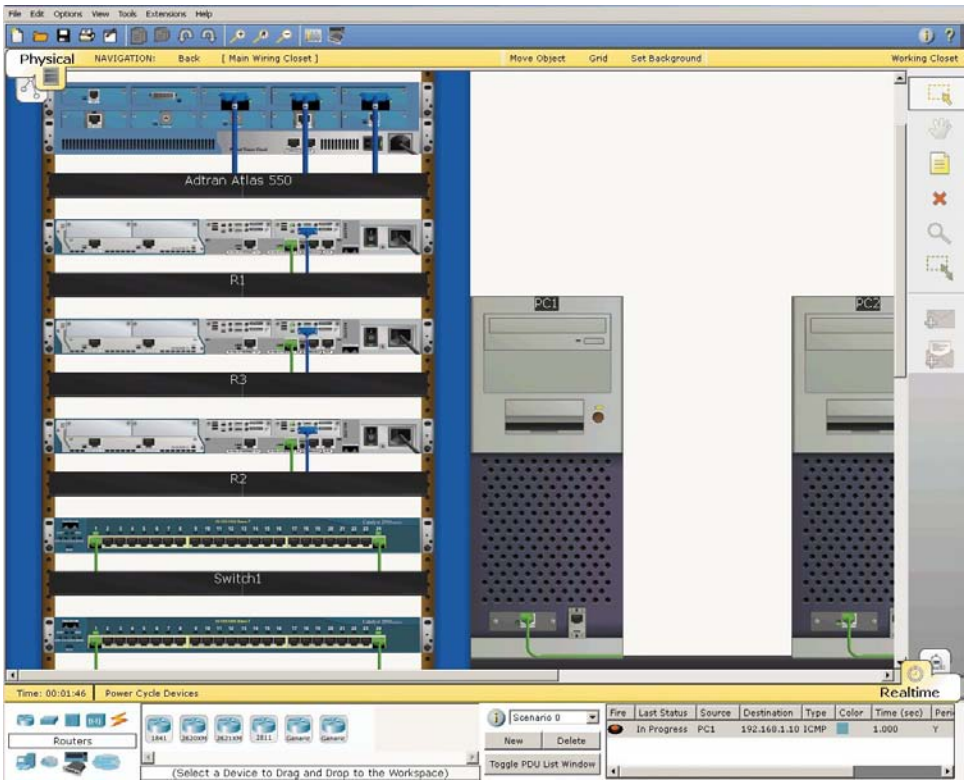

Figure 6. Screenshot of Packet Tracer, showing the representation of router configuration.

A second set of examples comes from work on simulation-based assessments for science in traditional schooling (Quellmalz et al., 2011). Figure 7 contains a screenshot from a simulation-based task targeting student understandings of organism classifications and food webs. In this task, students view animated sequences of organisms interacting in the ecosystem and are asked to draw food webs in dynamic environments. Figure 8 contains a screenshot of a related task targeting understanding of ecosystems, in which students dynamically interact with features of the ecological system to monitor population patterns to achieve desired ends such as an environment that is stable in terms of the populations of various organisms. Tools in this environment include slider bars to manipulate features of populations, a visual depiction of population density, and time series line plots of the populations of organisms over time. These tools are coupled with questions prompting students to make predictions, examine them, and indicate their conclusions and justifications. Video demonstrations of these and a number of other simulation-based tasks may be found online (SRI International, 2007).

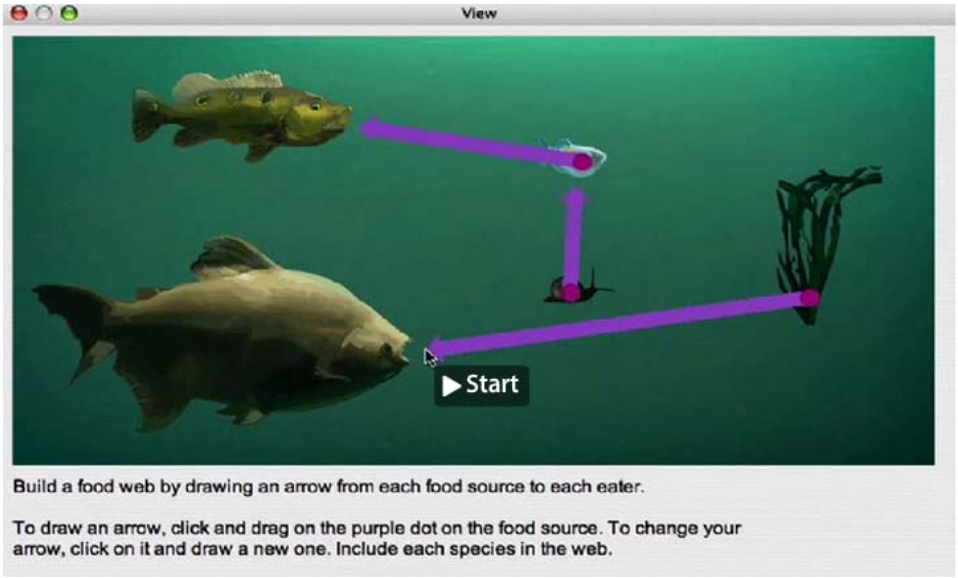

Figure 7. Screenshot of organism and food web simulation-based assessment.

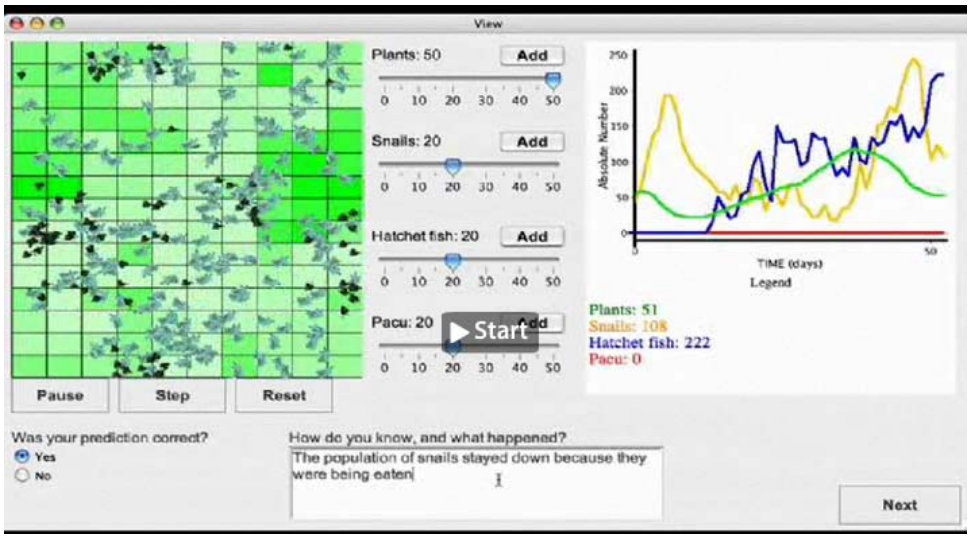

Figure 8. Screenshot of ecosystem simulation-based assessment.

Psychometrics of Simulation-Based Assessments

With these sorts of tasks, what in our psychometrics and measurement modeling has to change? Actually, nothing *has* to change. If our inferential goal is (still) to order students on one broadly construed conception of proficiency represented by one SMV, and we think of evidence from student performance in terms of a single OV for each task, and we deem these performances and resulting OVs conditionally independent given the single SMV, then very little, if anything, has to change. For example, if we think of troubleshooting a computer network in a Packet Tracer task in terms of a single proficiency ("troubleshooting computer networks"), and characterize each student performance in terms of one observable in two categories (e.g., the network is fixed or is not fixed), and think that the performance on such troubleshooting tasks is conditionally independent given the single SMV, modeling the OVs via a 2PL may indeed be appropriate.

However, this is less than desirable for a few reasons. First, simulations may be a very inefficient way to obtain this sort of evidence. Packet Tracer tasks may be so involved that students take upward of a few hours to complete them (much like real-world tasks in computer networking). A couple of hours' work to yield one OV seems like a wasted opportunity, and indeed, it may be unrealistic to then expect students to complete many such tasks for an assessment, as we do with shorter, more disconnected tasks in traditional assessment.

More importantly, interpreting and modeling evidence from simulations in such a limited way misses out on the real evidentiary opportunities that often motivate the use of simulations. Simulations are appealing because they afford us the potential to measure multiple aspects of proficiency and offer a multitude of evidence in ways not easily accomplished in traditional assessment formats. In Packet Tracer, we might conceive of the relevant proficiencies in terms of working with different devices (e.g., routers, switches, computers) and/or in terms of various actions that occur across devices (e.g., configuring passwords, allowing connections). Or we might wish to characterize proficiency in terms of both accuracy and efficiency and look for different though related data to constitute evidence, such as whether a feature of the network is fixed and how long it took to fix it, respectively.

Thus, in the balance of this report, when we refer to simulation-based tasks or assessments, we not only mean these kinds of tasks but the *kinds of evidentiary reasoning facilitated by the use of these tasks*. We will turn now to characterizing a number of features of these situations in terms of the elements of the CAF. Along the way, we will present key principles and challenges of evidentiary reasoning in assessment and illustrate them with examples from traditional assessment environments. The examples are purposefully selected as occurring *outside* simulation-based assessments to illustrate the generality of these principles and their importance to psychometrics in any context. The resolutions of the issues in traditional assessment are rarely made explicit and/or are often taken for granted as a done deal. This likely reflects the organic as opposed to systematic development of assessment practice over time, which occurred against a backdrop of (a) often unstated background beliefs regarding psychology, learning, and expertise (Mislevy, 2006); (b) certain purposes of assessment; and (c) technological constraints of various sorts, including not only the delivery of assessments but also machinery for storing, manipulating, and analyzing data from assessments (DiCerbo & Behrens, 2012). Importantly, the challenges posed by these principles become exacerbated in simulation-based assessments, and we will see that either (a) they cannot be resolved in the usual ways, or, if they can, (b) we often do not want to do so, lest we miss out on realizing the potential of simulation-based

assessment. Put another way, the opportunities for more intricate reasoning that are afforded by simulation environments coupled with advances of technology allow us—if not force us—to address these issues in ways that are markedly different from the commonly assumed backdrop prevalent in traditional assessments.

Task Model

Principle 1: What is presented to the student matters crucially for evidentiary reasoning, in ways we might not intend or expect. Figure 9 presents alternative versions of a question targeting whether the student knows the capital of Australia. Consider the differential evidentiary value of a correct answer (Canberra) in each of these situations. Suffice it to say that we might argue whether or not these items are of different difficulty, or we might debate if they even measure the same construct, or whether they are appropriate for, say, inferences about a student's knowledge of geography. We might adopt the position that there are not really any differences other than one of relative difficulty. Or we might argue that not only are they of different difficulty but also they provide evidence about different things; in contrast to Version a, Version b tells us something about the student's knowledge in the context of Australian national and territorial capitals, Version c tells us about his or her knowledge in terms of other world capitals, and Version d tells us nothing at all about the student's knowledge of Australian national or territorial geography or any other nations' capitals.

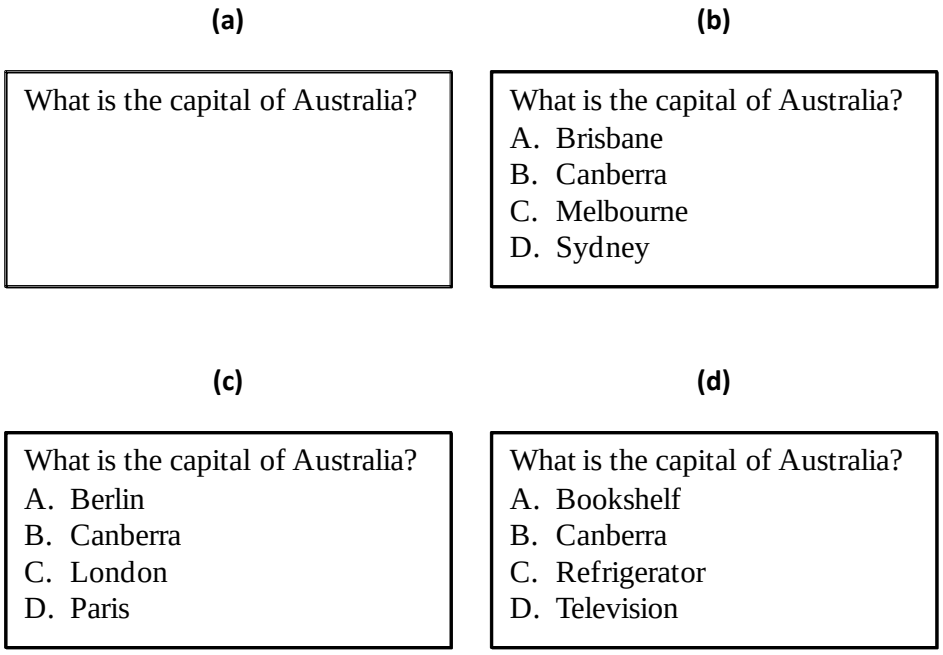

Figure 9. Four versions of a question about the capital of Australia.

Recognizing the sensitivity of the evidentiary quality to features of the task, note that the difference between traditional and simulation-based assessments is perhaps most stark in the tasks themselves—what is presented to the student, what the student interacts with, and the forms of the work products the student submits. Simulation-based assessments tend to have fewer tasks, perhaps even as few as one, that are each longer and more connected in that the actions students take are explicitly dependent on one another. Computer-based delivery also supports the recording of more complex work products. For example, work products of tasks in Packet Tracer could include the final configuration of the network, time-stamped log files of what was typed into the command line interface, and click streams of their use of the system.

Simulation tasks might also be structured or scaffolded in ways not present in static assessments. A popular example of the structuring is facilitated by the unfolding nature of the task and the student–system interactivity. For example, how a student configures a password on a device might impact what else may need to be done. This is particularly salient if the assessment involves multiple attempts. For example, suppose a student working through the ecosystem task depicted in Figure 8 set the levels of the populations in such a way that one organism died out. Having the student engage with the task again represents a structuring based on the student's past behavior. A similar structuring occurs in game-based assessments in which students repeat a game level until they complete it successfully (e.g., Kerr, Chung, & Iseli, 2011) and assessments in intelligent tutoring systems that provide repeated attempts, with or without hints as additional scaffolding (VanLehn, 2008; VanLehn & Niu, 2001).

Student Model

Principle 2: Assessments are almost always multidimensional. Despite the popularity of unidimensional measurement models, assessments are almost always multidimensional in the sense that students bring multiple distinct aspects of proficiency to bear when answering questions, solving problems, and completing tasks, especially in complex domains (cf. Reckase, 2009, on dimensionality from the perspective of the matrix of values of the OVs). This does not mean that the use of a single SMV and a unidimensional measurement model is necessarily unwarranted. But their use does imply certain simplifications or approximations, which may or may not be desired, depending on the purpose of the assessment. Unidimensional IRT models may be well suited for situations in which we desire the relative ordering of students in terms of a single dimension viewed as a coarse characterization of a truly multidimensional domain. Simple examples of managing the multidimensionality of the domain in support of a unidimensional student model include delimiting the relative weights afforded to different sets of tasks that target different aspects of proficiency, say, via tables of specifications, or content requirements/constraints in adaptive testing scenarios. In some circumstances, the multidimensional space can be ordered such that it collapses into a single dimension (Reckase, 2009). In other situations, especially when we wish to support inferences about multiple aspects of proficiency and such orderings of the multidimensional domain space are not substantively meaningful, a unidimensional model might not be sufficient.

In simulation-based assessments, the student model is often multidimensional in that we seek to make inferences about multiple, distinct aspects of the proficiency. This occurs for two reasons. First, the technology of task delivery and work product storage allows us to capture different forms of evidence. As a simple example, computer-based delivery and monitoring of student behavior allows for

the capturing of the time between actions as well as the actions themselves. This is also true of computer-based administrations of traditional task formats. This supports a distinction in the student model between accuracy and efficiency; see van der Linden (2007) for an example of such a model with two SMVs for a traditional assessment. In addition, one appeal of simulation-based tasks is their promise to provide evidence about multiple aspects of proficiency. One line of argument is that, as more authentic representations of real-world situations, simulation-based tasks allow for inferences about aspects of proficiency that are difficult to measure in traditional task formats that have less fidelity to real-world situations. Thus they offer opportunities to collect evidence about multiple aspects of proficiency, usually in finer grained delineations than the single SMV of traditional assessments, including the integration of those aspects of proficiency in tasks mirroring the complexity of real-world situations (e.g., Rupp et al., 2012).

Evidence Model: Evidence Identification Rules

Principle 3: The more open the workspace, the more possibilities there are—and the more decisions need to be made. Here workspace refers to the space of possible behaviors we might see, operationalized as the possible work products. The assessment community knows well how to handle Scantron sheets and other familiar forms of work products (e.g., written essays) because there is an explicitly defined—if not always articulated—clear association between behaviors and values of the OVs produced by the application of evidence identification rules to these work products. For example, a Scantron sheet associated with the question in Figure 9b might have the following rule associated with it:

> Evidence identification rule: Assign the OV the value of "1," standing for "correct," if the bubble next to option B is filled in; otherwise, assign the observable the value of "0," standing for "incorrect."

And even in situations where there is more ambiguity, such as ratings of constructed responses, systems have been developed to make them more manageable through defined rubrics and investigations of the reliability of the ratings through the use of multiple raters.

However, with evidence identification rules, we must account for all possible behaviors in the workspace. Figure 10 depicts several possibilities for what could be observed in a Scantron work product for the multiple choice task of Figure 9b, where "B" is the correct answer. Figures 10a and 10b represent easily interpretable work products to which we would assign correct and incorrect, respectively. Figure 10c is usually interpreted as a missing value (i.e., no answer submitted), and there are a number of common ways of interpreting this (e.g., as incorrect). The remaining panels in Figure 10 list but a few of the myriad possibilities we might see in the work product, for which decisions need to be made. Do we give credit for an incompletely filled-in bubble or the use of a different mark? What about answers written in on the side instead or in addition to bubbled-in answers? And what do we do when multiple bubbles are filled in? The point is not that there is a right or wrong way to handle these situations but that must have evidence identification rules for these and any other behaviors we might observe in the work product.

Despite the possibility of many behaviors, Scantron sheets are actually a fairly restrictive workspace. They are a knowledge representation by which students communicate their responses, and their very form helps to convey how the communication from student to assessor is expected to take place (Mislevy et al., 2010). Their use represents a choice on the part of the assessor regarding what he

or she will pay attention to and communicates as much to the student. A bit more casually, their use amounts to the assessor saying to the student, "Your response to the question will be judged in terms of your written record in this format, where a filled-in bubble corresponds to a selection of that option as the answer." Of course, clear directions and exposure of students to this format are important for them to be comfortable with this knowledge representation. Evaluating work products, then, comes to decisions about what to make of all the possible behaviors within this format (some of which are depicted in Figure 10).

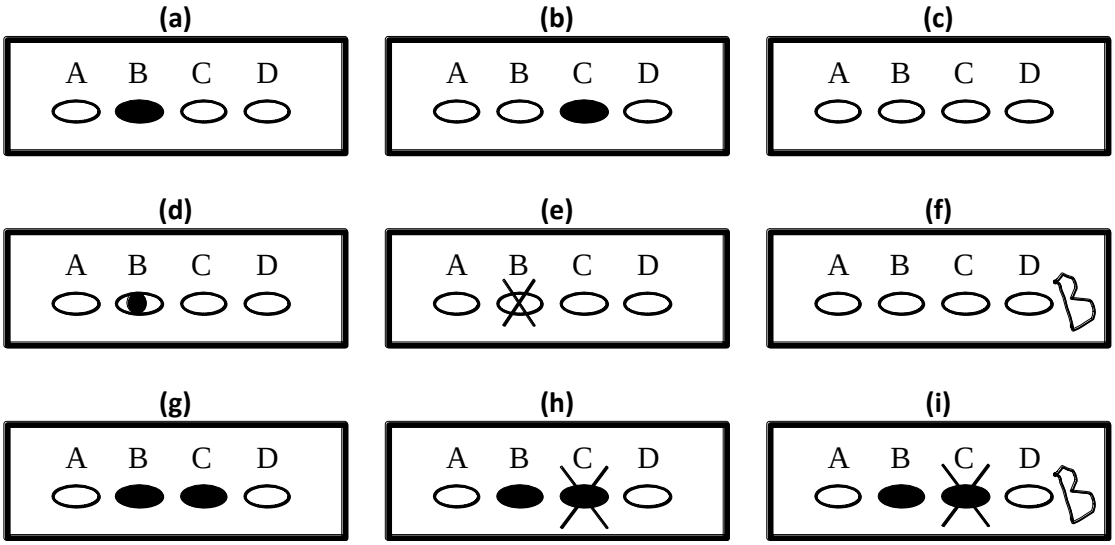

Figure 10. Example Scantron work products for a multiple choice question: (a) indicating a correct response; (b) indicating an incorrect response; (c) missing response; (d) incompletely filled-in bubble; (e) use of a different mark; (f) writing in the selection; (g) filling in multiple bubbles; (h) filling in multiple bubbles and an extra mark; (i) filling in multiple bubbles and an extra mark and writing in the selection.

Turning to simulations, if the simulation environment itself is new to a user, performance on tasks may be a function of the user's (un)familiarity with the environment. Moreover, the features of the environment undoubtedly affect the student's perceptions and interactions with the environment in ways we might yet not know. Reflecting on the research conducted and lessons learned in the last 100 years in reading assessment in terms of passage construction or selection, question writing, distracter creation and placement, instructions for filling in Scantron sheets, and so on, begs the question of whether a similar level of wisdom is needed about how to best design simulations. Such design choices are likely to be localized to the domain or task (e.g., how should we present arrows in food webs [Figure 5] or density and line graphs in ecological monitoring [Figure 6]?), though some generalities of design of simulation environments (e.g., Nelson & Erlandson, 2008) are likely to persist . Navigating the potential morass will be aided by advances in conceptualizations of design, such as Behrens, DiCerbo, and

Ferrara's (2012) characterization of the design of tasks in terms of the problem space, tool space, solution space, and response space.

The technology of simulations allows us to record and store much more than traditional assessment formats. A useful distinction is between the *end-of-state* work product (e.g., the final configuration of a computer network) and the *process* taken to arrive there (e.g., a log of all the commands entered to configure the computer network). Computer-based simulations allow for the recording of both. This runs us smack into the question of the rules for evidence identification: given that we can store so many features of student performance, what should we pay attention to? As discussed in the following subsection, the need to address this is, in the author's opinion, the single greatest challenge to simulation-based assessments.

The Openness of the Workspace and Grain Size

The source of the greatest potential for simulation-based assessments is also the source of its most daunting challenge. The openness of the workspace affords us the possibility to represent the real-world phenomenon of interest with a level of fidelity and authenticity that may not have been possible for economical, ethical, or other reasons. It simply is not feasible to offer every one of Cisco's students an unlimited supply of PCs, routers, switches, cables, and so on, on which to work though tasks, but with Packet Tracer, we can get remarkably close to the real thing. Nor is it feasible (or ethical) to have every middle school student go out to a lake and start making changes to the ecosystem to explore what happens to populations of organisms. Simulations afford us the opportunity to mimic these real-world phenomena. However, with this authenticity comes an openness of the workspace that allows for often an incredibly large and possibly infinitely number of behaviors that a student can conduct. In these wide-open environments, how, then, are we to know what to pay attention to that constitutes evidence? That is, how are we to define our evidence identification rules? In the following, we put forth several possible solution strategies.

First, we can limit our scope in some way. If the openness of the workspace is a problem, we can restrict it so that the space of possible behaviors is smaller. Another way is to "abstract up" what we pay attention to, from finer grained to more coarsely grained performance features. In thinking through all the sequences of actions that could be taken to successfully configure this computer network, we can abstract up to a coarser summary of performance. For example, in Packet Tracer tasks for troubleshooting computer networks, we might simply set up a rule analogous to that for the Scantron work product:

> Evidence identification rule: Assign the OV the value of "1," standing for "correct," if the packet can get from point A to point B; otherwise, assign the observable the value of "0," standing for "incorrect."

However, these strategies of restricting the workspace or what we pay attention to seem to have as their casualty the very things that are attractive about simulation-based assessments—the openness of the workspace, the possibility for students to behave as if in the real world, and the richness of student performance in terms of processes and nuances not captured by coarse summaries of performance. The use of the preceding evidence identification rule fails to aid in (a) informing on distinctions between, or characterizations of, the multitude of ways to fix the malfunctioning computer network, and (b) diagnosing strengths and weaknesses of the unsuccessful attempts. In short, the abstraction of the

information in a student's performance up to the level of correct–incorrect, as in the preceding evidence identification rule, is at too coarse a grain size.

Alternatively, we may avoid the problem of having too coarse of a grain size by enumerating all possible behaviors in the workspace and construct interpretations for all such possibilities. This is somewhat a function of the type of task and somewhat a function of its instantiation and the delivery in the actual assessment. Scantron sheets for multiple choice items are a workspace that affords many different possibilities, as in Figure 10. A different medium of presentation of the task and format of the work product, such as computer-based delivery, may eliminate some possibilities, though, of course, it may open up others.

Enumerating and forming interpretations for all possible behaviors may be feasible if the space of behaviors is fairly constrained or sufficiently scaffolded to rule out a number of otherwise possible behaviors. Of course, limiting the space of behaviors runs counter to the arguments of fidelity and authenticity of simulations. Accordingly, the workspace in simulation-based tasks is typically fairly more "open" in the sense that there are many, many behaviors in which a student can engage. In such contexts, we would be paralyzed if we had to specify all possible situations.

In laying out the evidentiary argument, we need to think through possibilities and make the evidentiary linkages as explicit as possible for all possible behaviors that are deemed important. Defining evidence identification rules is then a choice of the appropriate grain size, in which we navigate between having too coarse a grain size ("I will pay only attention to whether a packet can get from point A to point B") and too fine a grain size ("I will have an interpretation for each possible behavior I might observe"). The former is frustrating to those who seek to capitalize on the power of simulations; the latter is a prescription for possibly never completing the evidence identification rules. In practice, the answer lies somewhere in between. In our quest for more than just the coarsest characterizations of performance, choices need to be made about how to summarize or pool different behaviors into a more manageable subset. So the goal then becomes enumerating all possible behaviors or features of performance that have evidentiary bearing on the desired inferences and declaring the rest of these behaviors or features as irrelevant. We must declare when *a difference in performance* makes a *difference in our interpretation*. And we must *ignore differences where differences* they make *no difference*. If in configuring a computer network whether you used a router or a switch tells me something about you, we need to pay attention to that difference; if it does not, it can be can be safely ignored.

The assessment community has nearly a century of experience in making these choices for situations with fairly restrictive forms of work products (e.g., Scantron sheets), simple conceptions of proficiency (one SMV), and simple beliefs about the evidentiary bearing of performance on the tasks (corrector incorrect). For simulation-based tasks, in which the work products are in innovative formats and our conceptions of proficiency and the evidentiary bearing of performance are complex, the situation is more daunting.

Our solutions to this issue are likely to come from two sources. First, a principled approach to assessment design based on subject matter expertise facilitates the joint construction of tasks and evidence identification rules (and measurement models, discussed in the following section). During the design, we structure the tasks such that we know as much as possible about the features of possible behaviors in the workspace, which ones have evidentiary bearing, and which ones can be ignored.

Through design, we can set ourselves up to store and interpret relevant aspects of student performance, eliminate threats to interpretation, or streamline possibilities. For an example where the design of a highly structured workspace facilitates the specification of what features of performance to attend to, see VanLehn and Niu (2001) and Conati, Gertner, and VanLehn (2002). Saying we do this so that we know as much as possible about relevant features of possible behaviors reflects an acknowledgment that it is likely that we will not know all the relevant features of performance in complex tasks and open workspaces a priori, especially when there may be multiple ways to successfully complete the task and multiple ways to unsuccessfully complete the task.

A second source for knowing what to pay attention to comes from data analysis. Piloting the tasks and learning from data in various ways (cognitive labs, talk-alouds, larger deployments and calibrations) can offer us these insights. Analyses of data from administrating the tasks to students may reveal key features of performance to attend to in ways that yield new or revised evidence identification rules. In particular, the assessment community has much to gain from leveraging tools that have grown up in the educational data mining community for exploring and learning from data from simulation-based assessments (Mislevy, et al., in press).

In practice, the answer to the question of what to pay attention to is likely to be a mixture of all of these. If we required that we think of every possibility before rolling out an assessment for piloting, we might never complete the design of even one simulation-based task in an open workspace. The environment may be too complex for SMEs to anticipate all possibilities and articulate their evidentiary interpretations and relevance. Rather, a more constructive approach takes its cue from principles of modern statistical modeling and exploratory data analysis (Behrens, 1997; Behrens, DiCerbo, Yel, & Levy, in press; Box, 1976; Tukey, 1977) that recognize that any model is necessarily limited and interpret patterns of data relative to model-based expectations. This is done both to look to confirm those expectations and also to challenge them to find unanticipated patterns that reveal important features. In assessment, this unfolds as follows. We begin with principled design and construct the tasks, evidence identification rules, and measurement model in concert with desired inferences and what is believed about the domain. We then collect data from piloting, exploring patterns to support or refute the a priori expectations as well as to illuminate unanticipated features.

Typically, the activities in this latter stage of data analysis have focused on data–model fit between the measurement model and the data. Examples include statistical item analyses, checks for the dimensional structure, and analyses of noninvariance or differential functioning. The full implication of these practices can be seen by viewing the measurement model as the distillation of the assessment argument. An adequate data–model fit constitutes support for the evidentiary argument. A weak data–model fit points to weaknesses of the evidentiary argument, which may lead us to revise our measurement model, evidence identification rules, tasks and task models, inferential targets, or perhaps our understanding of the domain.

For examples where this interplay between principled design and data analysis have yielded improved understandings of features of performance in simulated-based assessments, see Rupp et al. (2012), in the context of understanding log files from Packet Tracer tasks, and Kerr and Chung (2012) and Kerr et al. (2011), in the context of identifying misconceptions and strategies in a simulation-based educational video game targeting rational number equivalence. These examples also highlight the nonlinear and iterative nature of assessment development; the sequencing implied by Figure 1

represents an idealized process of our intentions. In practice, things are much more iterative, with design followed by piloting and data analysis, which then informs on aspects of the domain farther back up the chain of ECD, leading to revisions of task, subsequent piloting and data analysis, and so on. We recognize that a complete a priori specification of all possibilities and interpretations is unlikely, and there is always a role for data analysis. Nevertheless, it is advanced here that negligence in design is likely to be costly. In the scale of effort for understanding what is going on and constructing evidentiary arguments in simulation-based assessments, experience teaches us that an ounce of design is worth a pound of data analysis.

Some other comments about the impact of a simulation workspace are warranted. There is usually a positive relationship between fidelity and openness and a negative relationship between openness and ease of interpretation. The real world is rich and complicated, and doing work in any domain involves lots of moving parts. The more authentic we wish to make a simulation, the more possible actions there will likely be, and the more will be required to specify the evidentiary argument.

In specifying the rules for evidence identification, the choices regarding what to pay attention to should be made with respect to the desired assessment goal. Formative uses may call for different aspects of performance to be monitored than summative uses. The desire to characterize students in terms of a profile of strengths and weaknesses might call for different foci than the desire to order students on a single coarsely conceived dimension.

The openness of the workspace is a challenge for identifying relevant aspects of end-of-state work products (e.g., the final configuration of a computer network). It is even more of a challenge for identifying relevant aspects of work products on the process (e.g., log files of all actions taken in configuring the computer network). Such features of performance have historically played a minor, if any, role in assessment, particularly at a large scale. Articulating the processes that constitute evidence regarding different categories, levels, or amounts of proficiency is arguably a much more daunting task than articulating the resulting end-of-state work products. There are often some general arguments to be made for end-of-state work products, such as students with higher levels of proficiency will tend to successfully complete tasks more often than students with lower levels of proficiency. It is much harder to articulate *how* they will complete these tasks. This is particularly so in complex situations, where there may be multiple distinct, possibly mutually exclusive processes that are equally valued. See DiCerbo, Liu, Rutstein, Choi, and Behrens (2011) and Rupp et al. (2012) for examples that focus on understanding such processes in Packet Tracer tasks.

In summary, the openness of the workspace is both a blessing and a curse. Indeed, the flexibility and fidelity of the workspace is often what makes it attractive for assessment. However, with that openness of the workspace comes the challenge of satisfying the principle that we need to know what to pay attention to, which amounts to a choice of grain size in our evidence identification rules.

Evidence Model: Measurement Model

Turning now to the measurement models, we discuss first what is really going on with measurement models, that is, how they embody the assessment argument and what they really require of us. We then discuss various aspects of how these requirements play out in simulation-based environments by describing a number of measurement model families that are well suited for the

simulation-based assessments. We then discusses two challenges and some possible solutions for the use of these models in simulation-based assessments.

Measurement Models as the Distillation of the Assessment Argument

Principle 4: We should be able to interpret any behavior we deem important; a measurement model articulates this by linking each possible value of an OV to each possible value of the SMV(s). Consider the evidentiary reasoning involved in the use of ordered multiple choice items (Briggs, Alonzo, Schwab, & Wilson, 2006). These multiple choice items are characterized by their options being linked to different, ordered levels of knowledge, proficiency, or expertise with respect to a domain, as in a learning progression. The key point for our purposes is that this approach makes explicit a connection between each possible response behavior and a particular interpretation. If a student responds with a certain option, it is interpreted as evidence that the student is at a particular level of the ordering. This evidentiary interpretation can be formally built into our psychometrics as follows. First, we define our evidence identification rules such that instead of parsing the student response into a dichotomous OV (for correct and incorrect), the OV will have as many levels as there are response options. And accordingly, we move from a dichotomous IRT model to a polytomous IRT model, specifically, one that capitalizes on the implied ordering to model each level of the now polytomous OV (see Briggs et al., 2006, for one such polytomous IRT model). In ECD terms, we are doing two key evidence model activities. First, we are defining our evidence identification rules, now drawing a distinction among a variety of behaviors and defining an observable with corresponding categories rather than collapsing those behaviors into the same category. That is, we assign to each distracter a different value of the OV rather than pooling all distracters together as "incorrect." Second, through the richer IRT model, we are defining our measurement model to cohere with (a) the OVs defined in the evidence identification rules and (b) the SMV they inform on.

More generally, what we need is to define the evidentiary bearing of all possible behaviors on the SMV of inferential interest. This is done in two stages. In the evidence identification rules, we define what features of performance we are going to pay attention to and how the features and distinctions we decide to pay attention to are operationalized as OVs. In the measurement model, we set up the relationship between the SMVs and OVs in terms of all the possible values for each.

It is crucial to unpack this last point. To do so, let us reexamine what is going on in Table 1 and how it is a distillation of the assessment argument. A measurement model such as that in Table 1 embodies much more than "Item *X* measures proficiency θ." It *requires* more, and it *does* more. It requires as its inputs choices for the SMVs (here one SMV specified with two possible values) and the evidence identification rules to produce OVs (here one OV with two possible values). And what it does is specify the connection between the SMV and OV by specifying the probability *for each possible value of the OV conditional on each possible value of the SMV*. By doing so, this embodies an assessment argument in that, for any possible value of the OV, we know what the evidentiary bearing is on the SMV. Returning to Table 1, if $X = 1$, our beliefs regarding θ are updated by using the likelihood of 9:2 in favor of the student being at a high level of θ. If $X = 0$, our beliefs regarding θ are updated by using the likelihood of 8:1 in favor of the student being at a low level of θ.

It is noteworthy that this goes on in all assessments, even if it is not recognized as such. Much headway has been made in large-scale assessment by specifying that there is one SMV that drives performance on tasks and that for each task, there is a single dichotomous OV based on correct–incorrect scoring. Finally, an IRT model such as that in (1) efficiently lays out the probability for any value of each OV conditional on any given value for the SMV by using only a few parameters. IRT models such as those in (1) may not appear similar to the measurement model in Table 1, but this is mainly because of the specification of the latent SMV as a continuous variable in IRT rather than as a discrete variable in Table 1. We can connect the two by viewing IRT as defining a continuum of possible values for θ and then structuring or smoothing the conditional distribution of X given θ in terms of a few parameters (i.e., a_j, b_j, and c_j in [1]). Importantly, what they share is that there is a conditional probability distribution for each value of the OV for each value of the latent SMV.

From this view, the measurement model is a distillation of assessment argument. It is the junction point between what a student does (the work product) and the inferential targets (the student model). It represents our understanding of the evidentiary bearing of student behavior, enacted in a statistical measurement model by laying out the conditional probabilities for each value of the OV(s) given each value of the SMV(s).

Shifting to simulation-based assessments, our measurement models do not *have* to change from our familiar forms. Familiar measurement models may be very suitable for simulation-based tasks, provided our purposes, constraints, and evidentiary arguments remain the same. In short, if we swap out our familiar tasks and insert in their place simulation-based tasks for which we can define evidence identification rules that yield OVs viewed as conditionally independent with respect to a single latent SMV, things will likely operate in much the same way as they do now. However, this will not be the case when we pursue simulation-based assessments in which

- performance on rich, authentic, integrative tasks often requires students to bring to bear expertise on multiple aspects of proficiency
- performance on such tasks can be characterized in terms of multiple, related features of evidentiary relevance

The first bulleted point suggests the specification of multiple SMVs. The upshot of the second bulleted point is that, for a single task, we may choose to pay attention to multiple aspects of performance and define multiple OVs accordingly. Importantly for the building of measurement models, this departs from the usual 1:1 relationship between tasks and OVs; that is, whereas the usual approach to a traditional assessment comprising 50 tasks (items) would specify 50 OVs (i.e., one for each task or item that summarizes correctness of performance), we recognize that multiple OVs can come from a single task.

As an aside, these considerations throw into sharp relief the reasons why tools like tables or graphics that "map" items to certain aspects of proficiency, claims, or SMVs are insufficient for laying out measurement models. But given that they have long been successfully used in assessment development, it is instructive to ask, Why are they so useful? Why are they so effective in developing assessments, including the development of the measurement model? To the extent that they are helpful, it is because there is an implicit or unstated argument. Simply saying "Item X maps to proficiency/skill/construct/claim/SMV θ and so is a good item to use in an assessment of θ" is actually

shorthand for a more complex argument, such as, "If we administer item *X*, and take all possible behaviors and distill them down into two categories, one for correct and one for incorrect, then having an observation in one of these categories tells me something about, or helps me make a desired distinction with respect to, θ." Tools like item maps are useful for communicating what are essentially approximations to a more complete evidentiary argument. As a basis for measurement models, they work fine as summaries in traditional assessment, but the gap between what they communicate and what is required is exacerbated as we pursue more complex evidentiary reasoning. With the desire to monitor multiple aspects of performance that come from the explicit construction of a task to yield such multiple evidentiary tokens, and the capability that comes from harnessing available technology, we can see that the usual ways in which we talk about assessments are insufficient. ECD provides a framework for couching the well-established practices of assessment in a general language of assessment that also supports moving beyond the (often unrecognized) boundaries of our traditional ways of discussing assessments. Our familiar tools and language serve us just fine for traditional assessments because they organically grew up at the same time as the purposes and constraints of these assessments. Once we want to operate outside of these bounds—as we want to when using simulations—we are well served to move to the more general language and representations of ECD.

The three following subsections briefly review popular and emerging families of multidimensional measurement models that are potentially fruitful for situations with multiple SMVs. Two succeeding subsections review issues surrounding key challenges to their implementation and strategies for addressing those challenges.

Multidimensional IRT

Natural extensions of the commonly used unidimensional IRT models are multidimensional IRT (MIRT) models, which specify OVs as dependent on multiple continuous latent SMVs. A useful distinction here is between OVs modeled as *factorially simple* or *factorially complex* (McDonald, 1999). An OV is factorially simple if it depends on exactly one latent SMV. Figure 11a depicts a model with all the OVs modeled as factorially simple. Each OV is at the destination of exactly one unidirectional arrow originating from a latent SMV, indicating the dependence of the OV on that SMV. The bidirectional arrow between the SMVs reflects the possibility of a correlation between them. In this case, the MIRT model takes on the appearance of patching together several unidimensional IRT models, and though a piecewise approach to modeling (i.e., fitting models for each latent variable separately) is possible, there are statistical advantages to a simultaneous analysis through MIRT (Zhang, 2004). An OV is factorially complex if it depends on multiple latent SMVs; Figure 11b depicts a situation where X_4 and X_5 are factorially complex. The statistical narrative for factorially complex OVs is that they should be modeled as conditional on *multiple* latent SMVs. This is appropriate when the aspect of performance captured by the OV depends on multiple aspects of proficiency.

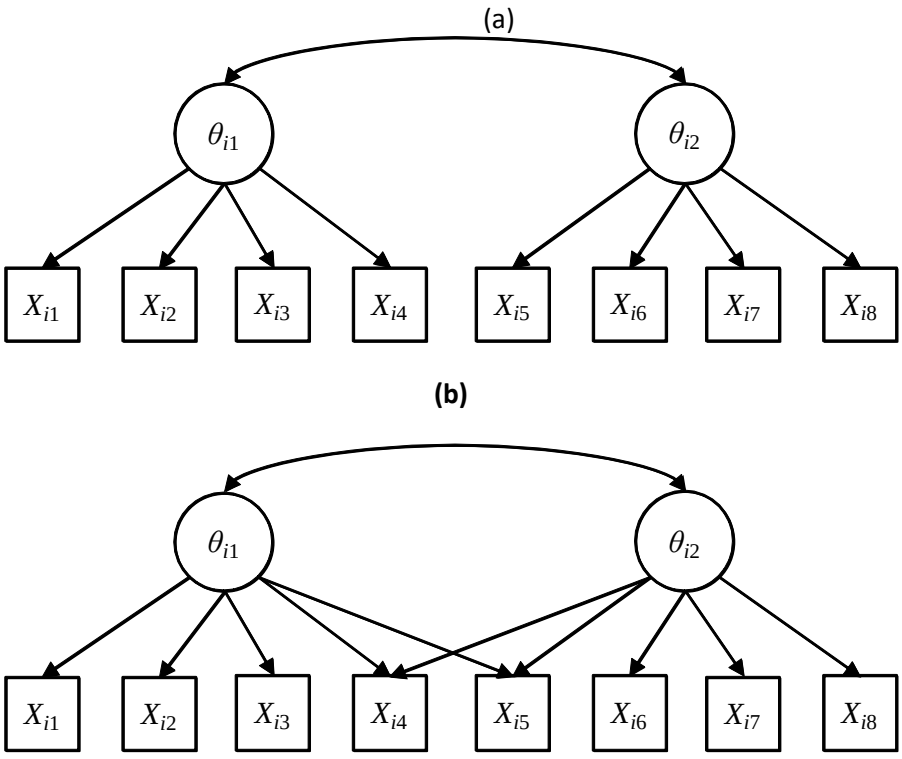

Figure 11. Graphical representation of a measurement model with (a) observable variables with factorially simple structure and (b) observable variables with factorially simple and factorially complex structure. Directed arrows indicate direct effects; bidirectional arrows indicate a correlation.

In specifying the model for factorially complex OVs, the analyst must also specify how the latent SMVs combine to produce or drive performance on that aspect of the task. The most popular choice are *compensatory* MIRT models (Reckase, 2009), which specify an additive function for combining the latent SMVs. For example, a logistic MIRT model for dichotomous OVs specifies the probability of an observed value of 1 (i.e., a correct response) for student *i* on OV *j* as (Reckase, 2009)

$$P(X_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j,c_j)=c_j+(1-c_j)\frac{\exp(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}{1+\exp(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)} \quad (2)$$

$$P(X_{ij}=0|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j,c_j)=1-P(X_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j,c_j),$$

where, in addition to the terms previously defined, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iM})'$ is a vector of *M* latent SMVs that characterize student *i*, $\mathbf{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jM})'$ is a vector of *M* coefficients for OV *j* that capture the discriminating power of the associated SMVs, and d_j is an intercept related to the marginal proportion of 1s (i.e., the difficulty of the task).

The use of a compensatory model reflects an assumption about the way the multiple aspects of proficiency combine in driving performance on the task. They are most appropriate in situations where the deficits along one aspect of proficiency may be compensated for by strengths of another (e.g., if a student's spatial reasoning proficiency may compensate for a relative lack of geometry skill in solving certain problems). This is operationalized by the summation in the exponent in the numerator of the first expression in (2).

Conjunctive MIRT models (Embretson, 1997) are alternatives to compensatory models that combine the dimensions via product terms:

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_i, \mathbf{b}_j, c_j) = c_j + (1 - c_j) \prod_{m=1}^{M} \frac{\exp(\theta_{im} - b_{jm})}{1 + \exp(\theta_{im} - b_{jm})}$$

$$P(X_{ij} = 0 \mid \boldsymbol{\theta}_i, \mathbf{b}_j, c_j) = 1 - P(X_{ij} = 1 \mid \boldsymbol{\theta}_i, \mathbf{b}_j, c_j), \tag{3}$$

where, in addition to the terms previously defined, $\mathbf{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jM})'$ is a vector of *M* location parameters for OV *j* that capture the location (difficulty) with respect to the associated SMVs. This model reflects a conjunctive structure through the product term in the first expression in (3) because a high probability of observing a value of 1 is obtained only if the student has high values on all the elements in $\boldsymbol{\theta}$ relative to the dimension-specific location parameter for the OV in $\mathbf{b}_j$. Such models may therefore have stronger connections to cognitive underpinnings of how students approach and solve tasks (Embretson, 1997).

Bayesian Networks

Bayesian networks (BNs; Jensen, 1996; Pearl, 1988) are a flexible family of statistical models that structure the joint distribution of variables via recursive conditional distributions. BNs employ discrete rather than continuous variables. In contrast to the MIRT models described previously, the SMVs in BNs are discrete latent variables. BNs may be represented as acyclic directed graphs (also referred to as directed acyclic graphs, DAGs), illustrated in Figure 12. On the surface, DAGs mimic those graphical representations presented earlier that follow common path analytic conventions, though certain technical distinctions exist (see Mulaik, 2009, chaps. 4 and 5). For our purposes, the key points about DAGs are that (a) a unidirectional arrow between variables indicates that the variable at the destination of the arrow, referred to as the *child*, probabilistically depends on the variable at the source of the arrow, referred to as the *parent*, and (b) DAGs are directed in the sense that the edges follow a "flow" of dependence in a single direction; in contrast to other graphical modeling traditions, the arrows are

always unidirectional rather than bidirectional. Thus, for each endogenous variable at the destination of an arrow, there is a probability distribution conditional on the variable(s) on which it depends. For each exogenous variable, there is an unconditional probability distribution.

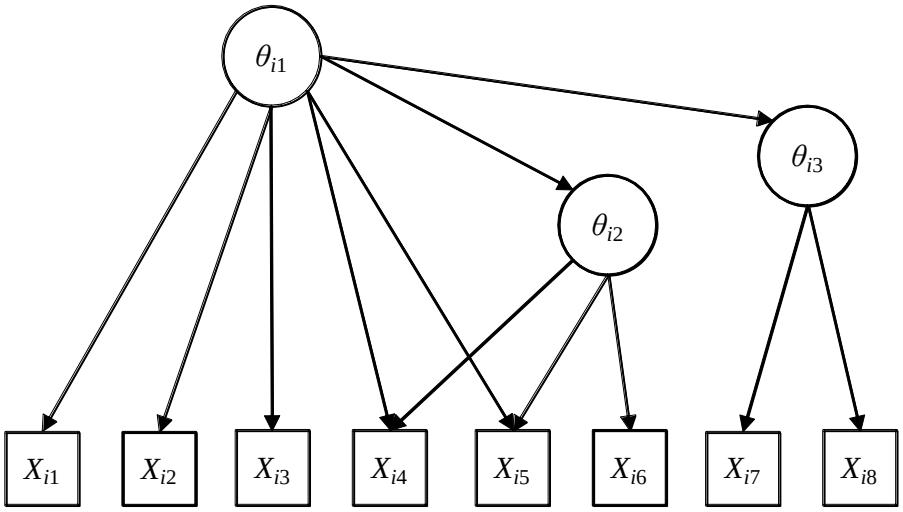

Figure 12. DAG for a BN measurement model.

The structure of the graph conveys how the model structures the joint distribution. Letting **Z** denote the full collection of variables $P(\mathbf{Z})$ may be factored according to the structure of the graph as

$$P(\mathbf{Z}) = \prod_{\mathbf{z}\in\mathbf{Z}} P[\mathbf{z} \mid pa(\mathbf{z})], \tag{4}$$

where $pa(\mathbf{z})$ stands for the parents of **z**; if **z** has no parents, $P[\mathbf{z} \mid pa(\mathbf{z})]$ is taken as the unconditional (marginal) distribution of **z**. Thus the graph reflects the dependence and (conditional) independence relationships in the model (Pearl, 1988).

For example, Table 1 illustrates a conditional probability distribution for an OV given a latent SMV. The conditional probability table for factorially complex OVs would expand on the structure in Table 1, with rows defining the combinations of the latent SMVs that are the parents of the OVs. Tables 2 and 3 contain examples of conditional probability structures, as may be specified for, say, X_4 in Figure 12.

Table 2. Example Conditional Probability Table for an Observable Variable Given Two Student Model Variables

Student model variables		Observable variable *X*	
θ_1	θ_2	1	0
High	High	.9	.1
High	Low	.6	.4
Low	High	.4	.6
Low	Low	.2	.8

Table 3. Example Conditional Probability Table for an Observable Variable Given Two Student Model Variables Reflecting a Conjunctive Relationship

Student model variables		Observable variable *X*	
θ_1	θ_2	1	0
High	High	.9	.1
High	Low	.2	.8
Low	High	.2	.8
Low	Low	.2	.8

As noted previously, BNs contain only unidirectional arrows. Hence, to model the joint distribution of the SMVs in Figure 12, we specify an unconditional probability distribution for the exogenous variable θ_1 and a conditional probability distribution for each of θ_2 and θ_2 given θ_1.

BNs are so named because they support the application of Bayes's theorem across complex networks by structuring the appropriate computations to yield posterior distributions for the unknown variables once data have been observed (Lauritzen & Spiegelhalter, 1988; Pearl, 1988). In the context of measurement models, once known values for the OVs are entered into the network, evidence accumulation occurs when the evidentiary import of the observed values on unknown variables is synthesized and propagated throughout the network (Mislevy, 1994).

BNs are a very flexible approach to building measurement models (Almond et al., in press). One can specify a variety of types of relationships, including additive and conjunctive relationships similar to those outlined in MIRT models as well as disjunctive and prerequisite relationships, and they support the specification of measurement models with dichotomous or polytomous OVs or latent SMVs (Almond, 2010; Almond et al., 2001; Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Levy & Mislevy, 2004; Mislevy et al., 2002). This flexibility supports the use of BNs across a variety of assessment applications, including those with complexities that pose difficulties for other methods such as longitudinal models for task performance and skill acquisition (Reye, 2004; VanLehn, 2008) and situations with serially dependent OVs as may be present in simulation-based assessments where multiple OVs are drawn from the same task, as discussed below in the section Managing Contextual Dependencies.

BNs can be employed with simulation-based assessments in situations where the data arrive and are modeled all at once (Almond et al., 2007; Levy & Mislevy, 2004). In addition, because BNs accurately and efficiently propagate the evidentiary impact of observations throughout the network once values for OVs are known, they support a modular approach to model construction and assembly. This supports dynamic assessment, including adaptive testing (Almond & Mislevy, 1999; Reye, 2004), intelligent tutoring systems based on simulations (Mislevy & Gitomer, 1996; Reye, 2004; VanLehn, 2008), and applications to simulation- and game-based assessments in which BNs are assembled on the fly as the situation evolves (Iseli, Koenig, Lee, & Wainess, 2010; Shute, Ventura, Bauer, & Zapata-Rivera, 2009; VanLehn, 2008).

Cognitive Diagnosis or Diagnostic Classification Models

A related family of measurement models is known as *cognitive diagnosis* or *diagnostic classification models* (DCMs; Rupp & Templin, 2008; Rupp et al., 2010). Many DCMs can be cast as BNs (Almond et al., 2007), and in this light, these DCMs can be viewed as BN models that reduce the parameterization of the conditional probability structure of the OVs via rules that specify how the latent SMVs combine. For example, a DCM that specifies a conjunctive rule states that for the OV to take on a value of 1 (representing a correct response to a task), the student must be at a certain level on each of the OV's parents. If the student has not reached the requisite level on one or more of these parents, he or she is not expected to have an OV value of 1 (i.e., are not expected to correctly complete the task). Unexpected correct responses are modeled via *guessing* parameters, and unexpected incorrect responses are modeled via *slipping* parameters, which may be specified at any of a few levels representing different theories of the response process (Rupp & Templin, 2008). Table 3 illustrates an example of a conjunctive rule where a correct response is expected if a student is at a high level of both SMVs, but the probability of slipping given that the student is at a high level of both is .10. If a student is at a low level of one or both SMVs, he or she is not expected to correctly respond to the task, but there is a .20 probability that the student will guess correctly.

Managing Contextual Dependencies

Principle 5: Understanding the evidentiary bearing of behavior might involve understanding other behavior; or, when synthesizing multiple observations, the whole may be more—or less—than the totality of its parts. We often need to synthesize the evidentiary import across observations in such a way that the whole is different—sometimes more, sometimes less—than simple aggregations afford. To illustrate this point, we elaborate on an example of the assessment of knowledge of Newtonian physics discussed by Braun and Mislevy (2005). Consider a two-item sequence regarding Newton's third law of motion, which states that for every action, there is an equal and opposite reaction. The first question poses a situation where a car and small truck of the same weight as the car are moving at the same speed and collide head-on. When asked about the relative amounts of force exerted by the truck and the car, the option "The truck exerts the same amount of force on the car as the car exerts on the truck" is correct and would constitute evidence of expertise with respect to Newton's third law. When

asked the second question, in which the small truck is replaced with a semi truck twice the weight of the car but traveling half as fast, the correct answer is the same. Students answering that the truck exerts a larger force on the car represents a misconception associated with thinking that the larger object exerts more force. Students answering that the car exerts a larger force on the truck represents a misconception associated with thinking that the faster object exerts more force. The point is that the interpretation and evidentiary import of a student's response to the second question depends on the first. If a student correctly answers the first question, then observing that the student answers the second question by choosing one of the incorrect options constitutes evidence of the student possessing the associated misconception. It might not have such an interpretation if the student had instead answered the first question incorrectly. From an evidentiary reasoning standpoint, interpreting the pattern of responses tells us more than considering them individually.

The reverse may happen in situations where responses are dependent on an additional, unmodeled source of covariation. For example, patterns of performance on tasks surrounding a common stimulus (e.g., multiple questions about a single passage in a reading assessment) might be due to ancillary features of the stimulus unrelated to the intended inference (e.g., the content of the reading passage). In these situations, failure to recognize these contextual effects in aggregations of OVs can lead to violations of the local independence assumptions, which may compromise the inferences or lead to overstating our precision about the inferences (Junker, 2010).

One approach to resolving these issues involves employing testlet models (Bradlow, Wainer, & Wang, 1999), which, when assuming compensatory relationships, may be viewed as special cases of compensatory MIRT models following a bifactor structure (Rijmen, 2010). Figure 13a depicts a bifactor representation of the model where the first OVs are formed from a testlet of tasks and the last four OVs are formed from a second testlet of tasks. Variable θ_1 is a SMV of inferential interest that influences all the OVs akin to θ in unidimensional models (Figure 4). Variable θ_2 is a SMV that serves to account for the associations among $X_1, \ldots, X_4$ due to those tasks functioning as a testlet; likewise, θ_3 for OVs $X_5, \ldots, X_8$. Conceptually, the use of this type of model aims at the appropriate evidence accumulation regarding the SMV of inferential interest by partitioning the sources of association among OVs. A related set of approaches specify a second-order latent variable model, where the first-order latent variables representing proficiency in particular contexts are modeled as a child of a second-order latent variable model representing proficiency more broadly construed. This model is depicted in Figure 13b. Examples of this approach can be found in IRT (Rijmen, 2010) and BNs for simulations (Conati et al., 2002; Mislevy & Gitomer, 1996). Theoretical work by Rijmen (2010) in the context of IRT has shown that the testlet and higher order models may be viewed as special cases of the bifactor model (Figure 13a), which offers flexibility for modeling relationships not supported by its more restricted versions.

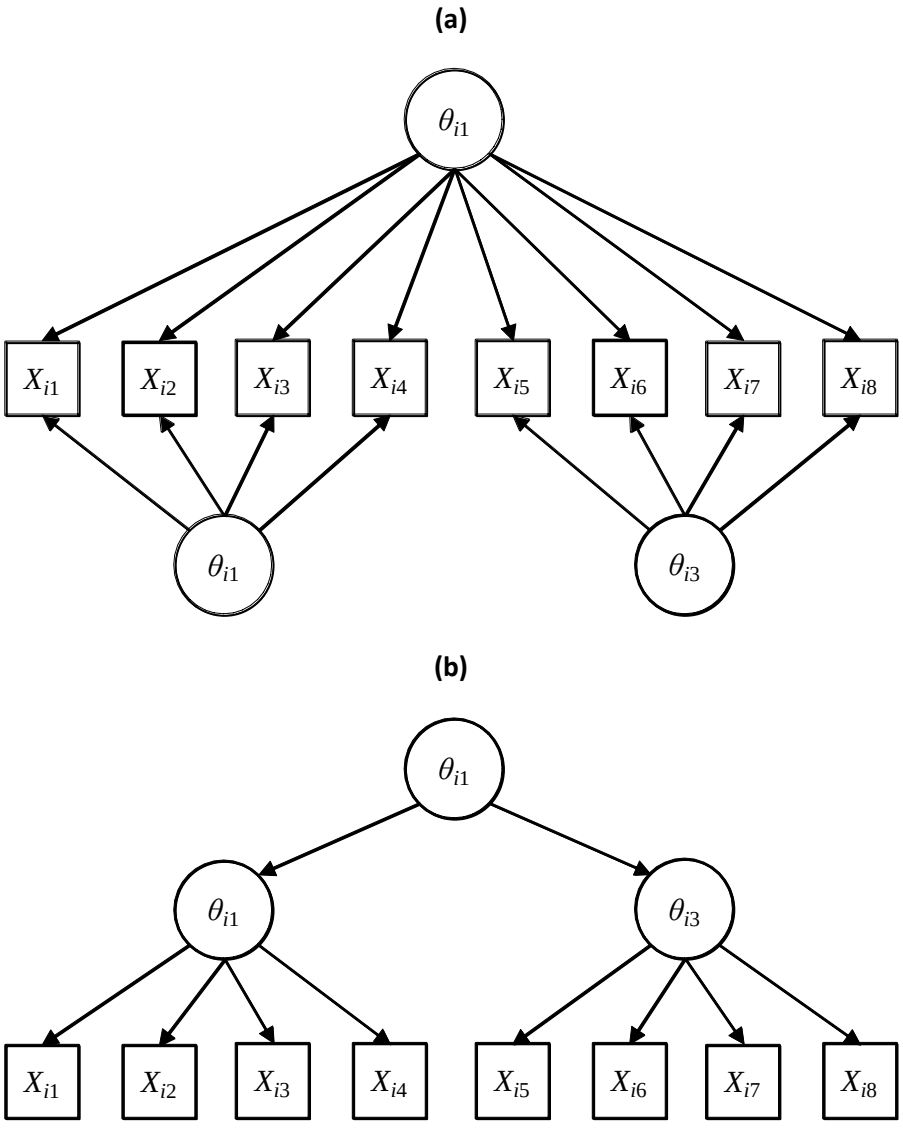

Figure 13. Graphical representation of a (a) bifactor measurement model and (b) second-order measurement model.

Testlet structures may also be modeled via conjunctive relationships. In a conjunctive IRT approach, the conditional probability of performance on a later task in the testlet is formulated as dependent on performance on an earlier task (Jannerone, 1997), as may be appropriate for situations where the interpretation of one performance depends on a previous performance, as in the Newtonian physics questions example. Almond, Mulder, Hemat, and Yan (2009) described a related approach that makes the conditioning of performance on one aspect of the task as dependent on another explicit by including directed edges between OVs. Figure 14 depicts such a model, where the arrow from X_1 to X_2

indicates that X_1 is a parent of X_2 and the conditional probability of X_2 is specified as depending on X_1 as well as θ. This can be extended in any of a number of ways; for example, X_5 , X_6, and X_7 in Figure 14 operate as a chain of such dependence dependencies. See Almond et al. (2009) for a description and evaluation of the use of compensatory, conjunctive, and several other types of relationships for managing the contextual effects and local dependence with BNs.

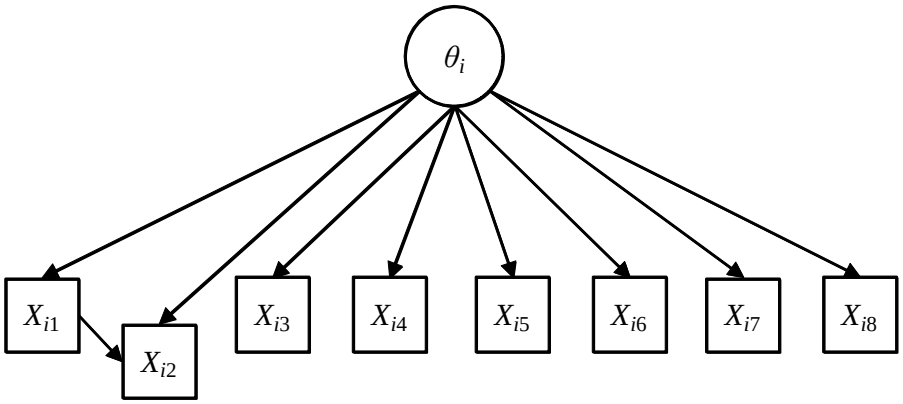

Figure 14. Graphical representation of a measurement model with a direct effect between observable variables.

From the perspective of the usual evidentiary argument in traditional assessment, these dependencies are something of a nuisance in that our interpretative narrative and measurement model built around disconnected tasks yielding conditionally independent OVs does not hold. But situations that give rise to these contextual dependencies in our data—namely, those in which what a student does at one point depends on what the student did previously—are likely to be the norm in complex simulation-based tasks. This may be due to explicit structuring of task or assessment. For example, in the ecosystems assessment depicted in Figure 8, the student is asked to make a prediction, engage with the simulation, evaluate the prediction in light of what the student saw, and then explain his or her thinking. This is even more prevalent in more open workspaces. In certain Packet Tracer tasks, the workspace (Figures 5 and 6) is so open that the sequence of actions is completely left to the student to determine. Strategies for managing these contextual dependencies in simulation-based assessments include specifying additional latent variables (Figure 13a) or directly modeling the dependencies (Figure 14); see Almond et al. (2009) and Levy and Mislevy (2004) for examples.

Where Do the Numbers Come From?

A challenge that is particularly acute for simulation-based assessments is that of the values of the conditional probabilities or the model parameters that govern them. In the context of unrestricted

BNs, this means the conditional probabilities that make up the conditional probability tables. There are a number of ways we might simplify this via parametric assumptions. IRT models may be used in ways that effectively smooth the conditional probabilities over the possibly many values of the SMVs (Almond et al., 2001; Almond et al., 2009; Levy & Mislevy, 2004). Similarly, DCMs simplify this process by starting with deterministic relationships (i.e., conditional probabilities of values 1 or 0) and then backing off such stringent assumptions with slip and guessing parameters, often constrained equally over OVs (Rupp & Templin, 2008).

Eventually, we will be left with the task of specifying the conditional probabilities, either directly or via a reduced set of parameters that govern them. There are two sources of information for those values. The first is via estimation from data. This follows the usual form as in other assessment contexts. We pilot the tasks, collect data, and estimate conditional probabilities or parameters accordingly. Compared to unidimensional IRT and CTT procedures, relatively little is known about sample size and related needs for estimating parameters to a sufficient precision in these more recently developed models. This strategy might be especially challenging for more involved simulation-based tasks that require longer times-on-task for students as well as yielding OVs that are not deemed conditionally independent. In traditional assessment formats, students attempt many tasks, and the resulting OVs are treated as conditionally independent. In simulation-based assessments, students might only engage in a few tasks, and to the extent that each interaction yields multiple OVs, it may not be appropriate to treat all the OVs as conditionally independent. What the sample size and piloting needs are in these contexts are not well established. This is further exacerbated in understanding features of performance corresponding to the processes in which students engage, as these tend to vary more than evaluations of the end-of-state in open workspaces. In some cases, there may be essentially an infinite number of behaviors in which a student can engage. If our evidence identification rules indicate that we must account for all of them and specify OVs with many values, then it is likely that no sample will be large enough to estimate all relevant conditional probabilities.

One approach for mitigating the needs of large samples involves leveraging collateral information. If evidence of proficiency can be gained from outside the assessment activity, that can be leveraged into the analysis of data from piloting. An ideal case would be if there was a mechanism for knowing each student's status on the SMV. We would then need to pilot the new tasks to a large enough sample of students at each level of proficiency to calculate conditional probabilities. Certainty regarding student proficiency represents a gold standard that can be approximated when perfect knowledge is unavailable. A well-established assessment of the target proficiency could be used to obtain estimates of SMVs, which could then be used in lieu of perfect knowledge of the gold standard. In this case, a Bayesian approach that yields posterior distributions rather than point estimates might be well suited to managing this uncertainty.

A second source of information is SME beliefs. We can set the values for the conditional probabilities or parameters that govern them based on SME expressions of things. For example, communications with SMEs might suggest that the probability that a student of high proficiency incorrectly completes a task yielding $X = 0$ is .1 and that the probability that a student of low proficiency correctly completes a task yielding $X = 1$ is .2. We would then set the values of the conditional probability table accordingly, as in Table 1. This has the advantage of being efficient, as no piloting, data collection, or estimation is required, but it has the disadvantage that it is entirely driven by a priori SME

opinion and, of course, may not accurately reflect the true relationships. Several simulation-based assessments have adopted this approach (Mislevy & Gitomer, 1996; VanLehn & Niu, 2001). Importantly, these assessments are of relatively low stakes, where the impact of less-than-the-best values for the conditional probabilities or their parameters is minimal. It is difficult to see how such an approach that does not make recourse to piloting and calibration could be used in high-stakes environments.

A third approach combines the information from SMEs with data analysis from piloting. A Bayesian approach to statistical modeling allows for the modeling of SME expectations and beliefs in the form of prior distributions for conditional probabilities or the model parameters that govern them. This can then be synthesized with data to yield posterior distributions. This is a powerful approach for leveraging all sources of information about complex assessments. For example, Levy and Mislevy (2004) detail the construction of BN for a simulation-based assessment in which the conditional probability tables are structured according to complex relationships and smoothed via IRT-type models, which greatly reduces the number of parameters. SME's expectations regarding the difficulty of the tasks are expressed via prior distributions, which are then combined with data from piloting to yield posterior distributions for the parameters and the resulting conditional probability tables.

Additional Challenges to Psychometrics for Simulations

The preceding sections have reviewed a number of recent advances in psychometrics and key challenges to them in need of attention as they are applied to simulation-based assessments. This section characterizes two other challenges.

Maturity of the Models

Over six decades of research and application, unidimensional IRT has matured to the point where there are well-known principles and procedures for addressing psychometric issues, including sample size needs for calibration and estimation, reliability/precision/information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data–model fit, differential functioning and invariance, and so on. As the preceding discussion hinted at, though MIRT, BNs, and DCMs are on solid footing statistically, they are in their relative infancy when it comes to their application as measurement models in larger assessment enterprises. Efforts are under way to tackle these issues within each of these modeling traditions; see Reckase (2009), Almond et al. (in press), and Rupp et al. (2010) for recent accounts of the states of these arts. What is needed is some basic psychometric research, both methodological and more applied, to increase our collective knowledge about the capabilities and limitations of these measurement modeling approaches.

Validation

Validation in simulation-based contexts poses a number of challenges. We might only have a few tasks, particularly if the tasks are involved and time consuming. Or we might be focusing on forms of data with which we are not accustomed to working, such as those derived from monitoring sequences of student actions. Or we might be using simulations because we think they afford us the opportunity to assess constructs that might be intractable with traditional formats (e.g., processes or sequences, how students respond to evolving situations, or efficiency). This poses the usual conundrum for developing new assessments: How much should the results relate to other assessments? If they correlate strongly, then are we really measuring something different? If they do not relate to other assessments, what,

then, could constitute evidence for our interpretations? How much is too much or not enough? A related set of issues surround questions of generalizability and transfer. To what extent are the interpretations made from a simulation-based assessment that contains perhaps only a few (more involved) tasks supported for instances in the domain that depart from those tasks or the simulation environment entirely?

Data analysis, including evaluations of data–model fit and associations with other variables, can support efforts toward validation. However it is advanced here that validation arguments are strongly enhanced by principled design. That is, rather than *hope for* validity, we should put our efforts into *building in* validity up front through the development of the assessment with the evidentiary argument explicitly in mind. The temptation to employ the technology at our disposal should be resisted until it is done in concert with an evidentiary assessment argument.

A Peek Into the Future

Short-Term Integration of Simulation-Based Tasks With Assessments

In the short term, the psychometrics of simulations are likely to look much like the psychometrics of traditional assessments. For assessment programs with existing CTT- or IRT-based measurement modeling, a first step would be to involve simulation-based tasks that conform to these measurement models and the evidentiary arguments they embody (i.e., short, disconnected tasks, with simple or abstract evidence identification rules yielding one or few OVs per task modeled as dependent on a single SMV). Over time, more complex simulation-based tasks and accompanying measurement models and evidentiary arguments can be developed, using strategies discussed next.

Model Building Versus Choosing

The preceding descriptions of MIRT, BN, and DCM models are suggestive of a set of choices regarding measurement models. However, much can be gained by viewing the situation less as one in which we *choose* an existing measurement model and more of one in which we *build* a measurement model for our specific needs. To see this, we briefly review recent research on statistical modeling in psychometrics that has led to key advances on two related fronts. First, connections among the various types of models are being explored. Examples include connections between factor analytic and compensatory MIRT models (McDonald, 1999), the placement of instances of each under broader frameworks (Mellenburgh, 1994; B. O. Muthén, 2002; Skrondal & Rabe-Hesketh, 2004), similar efforts couching DCMs in more general modeling frameworks (Henson, Templin, & Willse, 2009; von Davier, 2008), and their connections with BNs (Almond et al., 2007; Rupp & Templin, 2008). Similarly, models with discrete latent variables can be used to approximate those with continuous latent variables, and in some cases, vice versa; furthermore, in some cases, there are statistical equivalencies among them (Haertel, 1990). In short, the models share more than it may seem at first glance.

The second key advance concerns computation. Generally, improvements in estimation routines and software (Cai, 2010a, 2010b; Cai, Thissen, & du Toit, 2011; L. K. Muthén & Muthén, 1998–2010; Norsys Software Corporation, 2007; Rabe-Hesketh, Skrondal, & Pickles, 2004; Spiegelhalter, Thomas, Best, & Lunn, 2007; von Davier, 2005) have expanded our capabilities to fit complex measurement models. Among the many other developments, the last 15 years have seen the rapid rise of Markov chain Monte Carlo methods that have opened up a variety of possibilities for measurement models (see

Levy, 2009, for a review); particularly germane to the current focus is their application to complex BN models for simulation-based assessments (Almond et al., 2009; Levy & Mislevy, 2004).

With these modeling and computational developments, the psychometric community is capable of employing a wide variety of measurement models for complex assessment, including those that attend to features of simulation-based assessment. In particular, what the flexibility of these modeling and estimation paradigms provides is the capability to shift from a mode of *choosing* a measurement model to one of *building* a measurement model. The former is likely sufficient in simpler evidentiary reasoning contexts that constitute the majority of operational assessment. A 2PL IRT model may be perfectly sufficient if our evidentiary frame (a) targets a single broad conceptualization of performance along which we would like to differentiate students, taken as a cross-sectional snapshot and operationalized as a single SMV, and (b) contains many discrete tasks that each yield one dichotomous OV corresponding to correctness of the answer that can be treated as conditionally independent given the broad conceptualization of proficiency. To the extent that our evidentiary argument departs from this, we must modify the model. Slight departures call for slight model extensions. Simple examples include allowing for the possibility of guessing in selected response tasks via a 3PL structure for the resulting OVs, or using polytomous IRT models for polytomous OVs derived from finer grained characterization of performance, or managing a common source of dependence for certain OVs based on the task presentation via testlet model structures.

However, as we increasingly depart from this basic line of evidentiary reasoning, our measurement model will need to change accordingly. It has been argued here that the evidentiary narrative present in the use of simulations, involving

- multiple SMVs that may be related in complex ways
- multiple OVs from a single task, derived from features of both the end-of-state and the process by which it emerged
- interpretations of later behaviors that depend on earlier behaviors
- complex (e.g., conjunctive) relationships among the entities

may be so far afield from the usual narrative that rather than apply modifications to an off-the-shelf model, a better approach is to recognize that what we have is a wide tool kit of modeling components that can be assembled as needed to fit our purposes (Levy, Mislevy, & Behrens, 2011). Instead of asking "should I use a 2PL or 3PL here?" the operative question becomes "given the paradigmatic forms I have at my disposal, what should I build into my measurement model that is in concert with my student model, task model, and evidence identification rules?" That is, rather than choose a model, we build one using whatever components may be deemed necessary: got conditionally independent OVs?—familiar IRT models may be useful; have contextual effects?—use a testlet or multidimensional model; want to model multiple strategies?—mixture components may be folded in; want to also model efficiency or other aspects of proficiency?—specify additional latent SMVs; want to recognize clustering of students?—multilevel structures may layered. When viewing these statistical expressions as tools to call into service as needed, the landscape of possibilities of measurement models is greatly expanded. A modular approach to building measurement models also allows for the localized construction of model fragments, which can be specified and assembled as needed.

The payoff of these developments is that we have considerable more capabilities than is commonly believed and that the recently rapid pace of expansion is likely to continue. Thus building

assessments should be treated with an attitude that does not feel bound by beliefs about psychometric limitations. The recent conceptual and computational developments in statistical modeling have broadened the potential for measurement models to better reflect or rich theories. Viewing measurement models as narratives (Mislevy, Levy, Kroopnick, & Rutstein, 2008), we can tell better stories now—stories that are more nuanced and better aligned with the complexities of the real world and substantive theories about cognition, learning, and performance, which are potentially better represented in simulation-based environments.

Importantly, this means more tightly integrating psychometrics at the outset of the assessment development; that is, psychometrics for complex assessment arguments such as those that are invoked in simulations are *unlikely* to work when psychometric considerations are called into service long after design, delivery, and data collection. This approach may serve us adequately for assessments with simple formats, confined work products, and chunkable evidentiary arguments in the form of conditional independencies. But it is unlikely to work when we shift to more complex evidentiary arguments. Simulation environments with open workspaces that align with connected evidence structures are difficult to psychometrically tackle if those considerations are not incorporated into the design of the assessment and the argument it embodies.

Integrating Assessment With Learning and Instruction

Looking further ahead, simulations offer the potential to enact a number of changes to the typical way assessment is conducted. Simulations offer enormous potential for integrating assessment with learning and instruction; that is, we can replace the current "Teach. Stop. Test." mode in which the assessment is clearly marked as different from instruction and learning with one in which assessment is ubiquitously and seamlessly integrating with instruction and learning (Shute, Levy, Baker, Zapata, & Beck, 2009). See Behrens et al. (2008), DiCerbo and Behrens (2012), and Shute (2011) for examples and discussions of the use of simulations and related technologies for conducting such "stealth" (Shute, 2011) assessment in ways that are tightly integrated with instruction and learning. Packet Tracer, for example, was conceived of as a learning environment to be used for instruction rather than as a separate assessment environment. But just having the environment is not enough. What such integration will require, however, is a stronger alignment between substantive issues of knowledge and learning and performance in the domain with the psychometric model and the assessment.

If we continue with this vision of simulations as a way for assessment to become intertwined with instruction and learning, assessment then becomes less of a static, cross-sectional snapshot of proficiency and more of a longitudinal tracking of student proficiency as it changes over time. Importantly for psychometrics, most of our assessment arguments and measurement models are aligned with the "snapshot" view of assessment, in which it is assumed that learning during the assessment does not occur. This assumption will have to be discarded and techniques for modeling the change in student performance over time will need to be employed. Sources include developments in IRT (Embretson, 1991; Fischer, 1995, 2001; von Davier, Xu, & Carstensen, 2009), BNs (Reye, 2004), and general modeling frameworks drawing from structural equation modeling and multilevel modeling traditions (Skrondal & Rabe-Hesketh, 2004). Importantly, such tools may also be relevant in the context of a single assessment if the activity is one that facilitates learning, as in intelligent tutoring systems (Mislevy & Gitomer, 1996; Reye, 2004; VanLehn, 2008).

Flexibility in building measurement models will also bevitalas we increasingly need to build in contextualization to our assessment argument. As discussed previously, this will be important in simulation-based tasks with various levels of scaffolding. By extension, this will be even more important as we consider the role of assessment as embedded within other student activities surrounding learning. If the future of assessment includes mining the deluge from an ever-increasing digital ocean of information (DiCerbo & Behrens, 2012; Shute et al., 2009), our inferences and models will need to include recognition of the student's situation—what the student is working on, what the student has done in the past, and how to interpret the student's current behaviors in light of this.

Summary

Simulations pose opportunities for conducting innovative assessment, not only in terms of the student experience but also in terms of the assessment argument we can construct—what inferences we are able to make, what data we can collect, and how those data may be treated as evidence for facilitating those inferences. It is argued that a measurement model (a) may be viewed as a quantification of an evidentiary argument that seeks to reason from what students say, do, or produce to notions of their proficiency, and (b) is therefore inextricably linked with the adopted conceptions of proficiency, tasks deployed in service of assessment, and the interpretations of student performances on those tasks. Thus our usual approach to psychometric and measurement models may suffice if our evidentiary argument remains unchanged from its familiar form when employing simulation-based tasks. However, to the extent that our evidentiary argument changes, our measurement model will need to change as well.

Figures 15 and 16 compactly summarize the main theses regarding the core opportunities and challenges posed for psychometrics in simulation-based assessments as opposed to those in traditional assessment. Figure 15 lays out the inferential argument for traditional assessment as follows. We devise an assessment system in which we might see any of a number of possible behaviors. We seek to use these behaviors as grounds for the desired inferences. This is accomplished in evidence identification by processing the behaviors that occur in familiar formats that are relatively easily interpretable to produce OVs. These OVs are then entered into a unidimensional measurement model characterized by a single latent SMV, where this SMV is our representation of student proficiency used to make inferences and decisions about the student.

The key departures from this evidentiary narrative when we consider simulations are depicted in Figure 16. Most prominent is that the space of possible behavior in simulations is now much larger, and the desired inferences may be larger in scope or more nuanced as well. Behaviors are observed in possibly unfamiliar formats, and now it may be much more difficult to characterize salient aspects of these behaviors. The resulting OVs are modeled in much more complex ways, owing to their dependencies due to their contextualized nature and their hypothesized dependence on richer, multidimensional student models.

It is argued that recent developments have the psychometric community well poised to tackle these situations. In particular, modern data analytic methods coupled with principled assessment design offer a promising approach to meeting challenges and fulfilling the promise of simulation-based assessment.

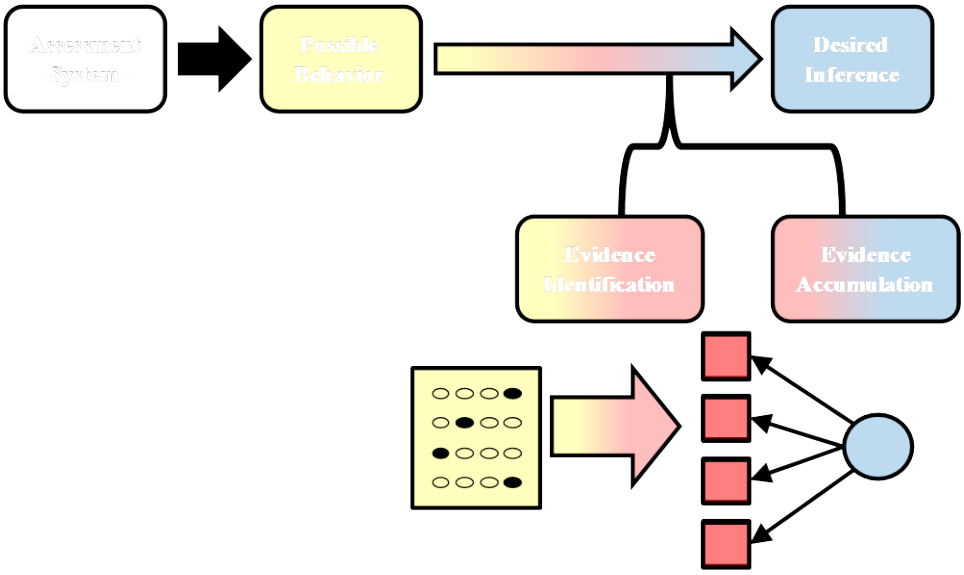

Figure 15. Schematic for the evidentiary argument for traditional assessment.

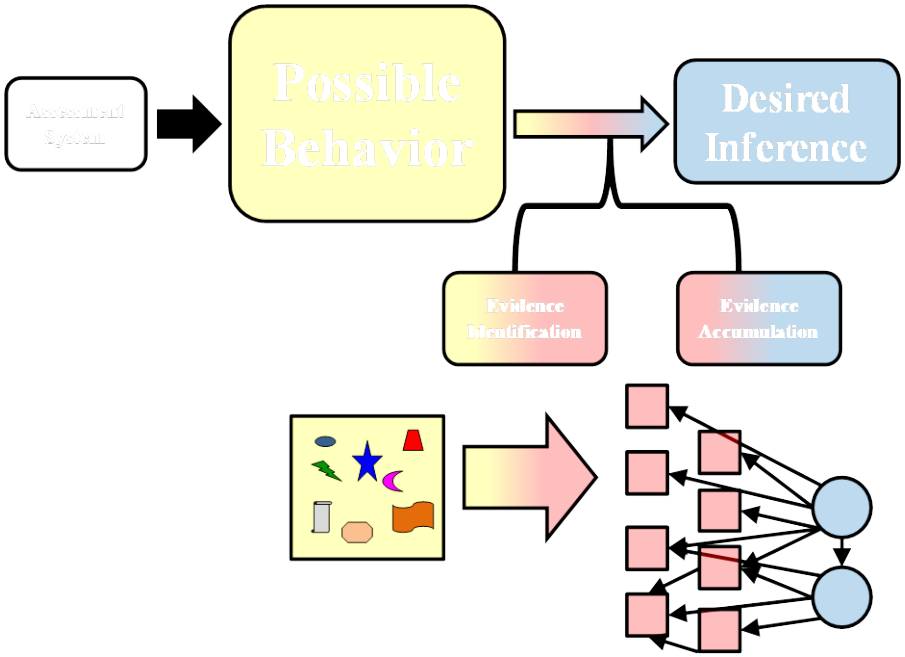

Figure 16. Schematic for the evidentiary argument for innovative simulation-based assessment.

Recommendations

We conclude with general recommendations about psychometrics for simulation-based assessments.

Recommendation 1

Simulation-based tasks may be most easily deployed or integrated into assessments in contexts where they are used in concert with traditional or existing evidentiary arguments. Simulation-based tasks may be used in support of evidentiary arguments that are richer than those of traditional assessments. Doing so will likely involve innovative measurement models.

Recommendation 2

The recent developments in statistical modeling allow for a broad set of choices for such innovative measurement models. Developers of simulation-based assessment will be well served to adopt a perspective that views that a measurement model can be built or customized to their specific needs. This will be best accomplished by including psychometric considerations from the outset and throughout the development of the assessment.

Recommendation 3

Basic psychometric research is needed on these innovative measurement models. Such research should address the use of measurement models in service of assessment needs and could include methodological or applied research on areas such as model calibration and parameter estimation, reliability, validity, test form creation, linking and equating, adaptive administration, data-model fit, and evaluating assumptions.

Recommendation 4

The key challenges to simulation-based assessment are likely to be most successfully addressed by integration of subject matter expertise and data analysis at all phases. Accordingly, psychometrics should play an integral role in the design, development, revision, and use of simulation-based assessments.

Acknowledgments

I wish to thank my current and past collaborators at Cisco, Pearson, ETS, the University of Maryland, WestED, CRESST, Arizona State University, and MW Productions, with whom I have had the pleasure of working on simulation-based assessments. I also wish to thank the Center for K-12 Assessment & Performance Management at ETS for their support.

References

Almond, R. G. (2010). "I can name that Bayesian network in two matrixes!" *International Journal of Approximate Reasoning, 51,* 167–178.

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., & Yan, D. (2001). Models for conditional probability tables in educational assessment. In T. J. Jaakkola & T. S. Richardson (Eds.), *Artificial intelligence and statistics 2001* (pp. 137–143). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44,* 341–359.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics, 34,* 491–521.

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).

Almond, R. G., Williamson, D. M., Mislevy, R. J., & Yan, D. (in press). *Bayes nets in educational assessment*. New York, NY: Springer.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131–160.

Behrens, J. T., DiCerbo, K. E., Behrens, J. T., DiCerbo, K. E., & Ferrara, S. (2012). *Intended and unintended deceptions in the use of simulations.* Princeton, NJ: Center for K-12 Assessment and Performance Management, Educational Testing Service.

Behrens, J. T., DiCerbo, K. E., Yel, N., & Levy, R. (in press). Exploratory data analysis. In W. F. Velicer & I. Winer (Eds.), *Handbook of psychology: Vol. II. Research methods in psychology* (2nd ed.). New York, NY: Wiley.

Behrens, J. T., Frezzo, D., 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. Wulfeck, & H. O'Neill (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). Mahwah, NJ: Lawrence Erlbaum. Associates.

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 (Eds.), *Technology-based assessments for 21st century skills*: *Theoretical and practical implications from modern research* (pp. 13–53). Charlotte, NC: Information Age.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71,* 791–799.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Braun, H. I., & Mislevy, R. J. (2005). Intuitive test theory. *Phi Delta Kappan, 86,* 488–497.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11,* 33–63.

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

Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–335.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cisco. (2010). Cisco Packet Tracer. Retrieved from http://www.cisco.com/E-Learning/prod/curriculum/cco_tdo_ldd/demos/PacketTracer/index.html

Cisco. (n.d.). *Cisco Networking Academy*. Retrieved from http://www.cisco.com/web/learning/netacad/index.html

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12*, 371–417.

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.

DiCerbo, K. E.,Liu, J., Ruststein, D. W., Choi, Y., & Behrens, J. T. (2011, April). *Visual analysis of sequential log data from complex performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dillon, G. F., Boulet, J. R., Hawkins, R. E., & Swanson, D. B. (2004). Simulations in the United States Medical Licensing Examination™ (USMLE™). *Quality and Safety in Health Care, 13*(Suppl. 1), 41–45.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495–515.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York, NY: Springer.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.

Fischer, G. H. (2001). Gain scores revisited under an IRT perspective. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Lecture notes in statistics: Vol. 157*. *Essays on item response theory* (pp. 43–68). New York, NY: Springer.

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.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika*, *55*, 477–494.Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic.

Henson, R. A., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74,* 191–210.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations* (CRESST Research Report No. 775). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf

Jannarone, R. J. (1997). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465–479). New York, NY: Springer.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York, NY: Springer.

Junker, B. (2010). Modeling hierarchy and dependence among task responses in educational data mining. In C. Romero, S. Ventura, S. Viola, M. Pechenizkiy, and R. Baker (Eds.), *Handbook of educational data mining* (pp. 143–155). Virginia Beach, VA: Chapman & Hall/CRC.

Kerr, D., & Chung, G. K. W. K. (2012). *Using cluster analysis to identify key features of student performance in educational video games and simulations*. Unpublished manuscript.

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Research Report No. 790). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R790.pdf

Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B, 50*, 157–224.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, *2009,* Article 537139. doi:10.1155/2009/537139

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4,* 333–369.

Levy, R., Mislevy, R. J., & Behrens, J. T. (2011). MCMC in educational research. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo*: *Methods and applications* (pp. 531–545). London, England: Chapman & Hall/CRC.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mellenburgh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. .J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Phoenix, AZ: Greenwood.

Mislevy, R.J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Research Report No. 800). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R800.pdf

Mislevy, R. J., Behrens, J. T., Bennett, R. E., DeMark, S. F., Frezzo, D. C., Levy, R., Robinson, D. H.... Winters, F. I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2). Retrieved from http://www.jtla.org/

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (in press). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction, 5,* 253–282.

Mislevy, R. J., & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 839–865). Amsterdam, Netherlands: North-Holland/Elsevier.

Mislevy, R. J., Levy, R., Kroopnick, M., & Rutstein, D. (2008). Evidentiary foundations of mixture item response theory models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 149–175). Charlotte, NC: Information Age.

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., & Yan, D. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Technical Report No. 580). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: CRC Press.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Nelson, B., & Erlandson, B. (2008). Managing cognitive load in educational multi-user virtual environments: Reflection on design practice. *Educational Technology Research and Development, 56,* 619-641.

Norsys Software Corporation. (2007). *Netica manual.* Retrieved from http://www.norsys.com/netica.html

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Kaufmann.

Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2011). 21st century dynamic assessment. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills*: *Theoretical and practical implications from modern research* (pp. 55–89). Charlotte, NC: Information Age.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual* (2nd ed., Working Paper No. 160). Berkeley: University of California Division of Biostatistics. Retrieved from http://www.bepress.com/ucbbiostat/paper160

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reye, J. (2004). Student modeling based on belief networks. *International Journal of Artificial Intelligence in Education, 14*, 1–33.

Rijmen, F. (2010). Formal relations and empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372.

Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S., Crawford, A. V., Calico, T., Benson, M. ---Behrens, J. T. (2012). *Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment*. Unpublished manuscript.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.

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

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. 503–523). Charlotte, NC: Information Age.

Shute, V. J., Levy, R., Baker, R., Zapata, D., & Beck, J. (2009). Assessment and learning in intelligent educational systems: A peek into the future. In S. D. Craig & D. Dicheva (Eds.), *Proceedings of the Artificial Intelligence and Education (AIED 2009) Workshop on Intelligent Educational Games,* Volume 3: Intelligent educational games (pp. 99–108). Brighton, UK: International Artificial Intelligence in Education Scociety.

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. Vorder (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2007). *Winbugs user manual: Version 1.4.3*. Cambridge, England: MRC Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

SRI International. (2007). *Calipers: Simulation-based assessments.* Retrieved from http://www.wested.org/calipers/assessments.html

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). New York, NY: Erlbaum.

VanLehn, K., & Niu, Z. (2001). Bayesian student modeling, user interfaces, and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education, 12,* 154–184.

Vargas, T. (2012, April 7). Priests in training confess their fears before stepping into the booth. *The Washington Post*. Retrieved from http://www.washingtonpost.com/

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61,* 287–307.

von Davier, M., Xu, X., & Carstensen, C. H. (2009). *Using the general diagnostic model for measuring learning and change in a longitudinal large scale assessment* (ETS Research Report No. RR-09-28). Princeton, NJ: ETS.

Zhang, J. (2004). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (ETS Research Report No. RR-04-44). Princeton, NJ: ETS.

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.