


SOCIOCOGNITIVE AND ARGUMENTATION PERSPECTIVES ON PSYCHOMETRIC MODELING IN EDUCATIONAL ASSESSMENT

ROBERT J. MISLEVY 

UNIVERSITY OF MARYLAND

Rapid advances in psychology and technology open opportunities and present challenges beyond familiar forms of educational assessment and measurement. Viewing assessment through the perspectives of complex adaptive sociocognitive systems and argumentation helps us extend the concepts and methods of educational measurement to new forms of assessment, such as those involving interaction in simulation environments and automated evaluation of performances. I summarize key ideas for doing so and point to the roles of measurement models and their relation to sociocognitive systems and assessment arguments. A game-based learning assessment *SimCityEDU: Pollution Challenge!* is used to illustrate ideas.

Key words: assessment argument, automated scoring, measurement models, interactive tasks, sociocognitive perspective.

1. Introduction

Significant developments are taking place in educational assessment and educational measurement: in technology, for assessments that can be interactive, immersive, simulation-based, created, and personalized, on the fly; analytic methods, in modeling and computation, learning analytics, machine learning, and natural language processing; systems and strategies to better integrate assessment with learning; and interest in higher-order capabilities, such as systems thinking and collaboration. I focus here on two foundational advances that help us put these developments to work effectively and validly, as they connect with longstanding concepts and principles from psychometrics. The advances are these:

- A sociocognitive psychological perspective, which concerns how people develop capabilities and use them to interact in the social and physical world.
- Assessment argument structuring, which explicates issues in design and inference in ways that a measurement paradigm alone does not.

I draw on three projects I have recently been involved with: *Sociocognitive foundations of educational measurement* (Mislevy, 2018) expands further on these two themes. The chapters of the edited volume *Computational psychometrics: New methodologies for a new generation of digital learning and assessment* (Von Davier et al., 2021) go more deeply into new analytic methods for measurement modeling and data analytics from this perspective. The *Handbook of automated scoring: Theory into practice* (Yan et al., 2020) provides further theory and examples on the evaluation of complex performances, connecting the concepts and methods of educational measurement with concepts and methods from data analytics and language processing to evaluate complex performances.

Correspondence should be made to Robert J. Mislevy, University of Maryland, Annapolis, MD21409, USA.
Email: rmislevy@umd.edu

1.1. Connecting Psychological Processes and Assessment Arguments

Samuel Messick (1994) proposed a way to begin thinking about assessment design:

A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute 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)

Some notion of the nature of capabilities and how people develop them and use them is implicit in this quote. This is where a psychological foundation comes in. Note also that a construct construed in this way is historically and socioculturally located. It concerns regularities in behavior among people and situations, in some milieu of activity, as relevant to the purpose of the assessment. In any application, assessment designers need to determine how Messick's general elements play out in the context at issue. Note finally that the quote is the core of an assessment argument, supporting both assessment design and score interpretation. We will see that there is more to it.

1.2. An Example: *SimCityEDU: Pollution Challenge!*

While the conception I describe also applies to familiar forms of educational assessment such as multiple-choice items and written responses, I mean to highlight new forms, often digital. The following projects illustrate aspects of the ideas:

- Unobtrusive modeling of students' proficiencies (aka "stealth assessment") in computer game-based assessment with theory-based task design and a Bayes net student model updated by automated evaluations of students' solutions (Ke & Shute, 2015).
- A scenario-based science assessment in which students work through menu-based branching conversations to predict the likelihood of a thunderstorm (Liu et al., 2016).
- A simulated science lab with theory-based investigations, affordances, and automated scoring procedures that generalize across tasks (Gobert et al., 2012).
- A tutoring system that automates evaluation, adapts problems, and provides interactive feedback for Newtonian physics, with an underlying measurement model to manage evidence and inference (Conati et al., 2002).

In this article, I will refer to a game-based assessment called *SimCityEDU: Pollution Challenge!* (Mislevy et al., 2014). *SimCityEDU* is a formative assessment, focused on systems thinking, embodied in a series of challenges in an environment based on the *SimCity* commercial game. It is designed around the more general five-level learning progression for systems thinking shown in Table 1, adapted from Brown (2005) and Cheng et al. (2010). Figures 1 and 2 are screenshots from one of its challenges, Jackson City, designed to provide an experience with an energy/pollution system that requires level 4 thinking to solve a problem and to educe evidence about a student's thinking through their actions.

It is already apparent that its design reflects the Messick quote. The levels of the learning progression jointly bring in (1) the capabilities of a person in terms of thinking about a system, (2) key features of a system that underlie a situation involving that system, and (3) actions a person can take in interacting with the system. These relationships drive the technical aspects of evidence identification and measurement modeling.

SimCityEDU does not consist of prepackaged items with easily identifiable and scorable responses. For example, a student tackling the Jackson City challenge to reduce pollution while maintaining electric power and commerce uses tools to explore the city, come to understand the problem, interact with the city through zoning and building actions, and examine their effects.

TABLE 1.
The systems-thinking learning progression.

Level	Competency level Description
1	<i>Students have a fragmented understanding of aspects of systems.</i> They may have partial knowledge of some of the definitions of system terms but cannot use them in a consistent or strongly coherent manner. While they can identify outcome variables (e.g., stocks that are explicitly part of the goal state), they are not able to track a causal link and they largely focus on macro-level directly observable variables. Their predictions and explanations are acausal, i.e., more assertions than cause-and-effect relations (e.g. “things happen because that’s the way they are.” Brown, 2005, p. 7)
2	<i>Students have an elemental understanding</i> (Brown, 2005, p. 7) of some aspects of systems—they can use models to represent simple, single cause-and-effect relations but without strong justification, i.e., they are still prone to common misconceptions, e.g., they tend to only relate macro-level, directly observable causes and effects rather than identifying hidden variables and factors. This is due in part to not being able to understand and analyze a system at different levels (Cheng et al., 2010). They are better at explaining than predicting
3	<i>Students have a locally coherent understanding of many aspects of systems.</i> Students can use system thinking terms to describe components and system relations in some contexts and use different representations. They can use models to represent bivariate cause-and-effect relations along with strong justifications. They can relate binary combinations of hidden and directly observable combinations, and even single causes to multiple effects. They are less prone to common misconceptions but still are limited to linear thinking with single causes (which may or may not be chained together.) They have a rudimentary understanding of negative feedback and can use it to explain and predict changes in the behavior of a system over time. They still are not able to consistently understand and analyze a system at different levels (Cheng et al., 2010)
4	<i>Students can relate multiple causes to multiple effects</i> as long as they behave in simple ruleful ways (e.g. cases in which all causes are needed for the effect to occur, cases in which all causes contribute independently to the amount of the effect as in Jackson City, etc.; that is, the causes are not emergent but are explainable in terms of the causal component parts.) This level is consistent with Brown’s (2005) conceptual depth level 4. Students can apply this scope of understanding within a wider range of contexts than in prior levels
5	<i>Students have a globally coherent understanding of many aspects of systems thinking in many contexts.</i> They can analyze of moderately complex system that includes multiple variables that may include hidden variables, feedback spread out in space and time, and emergent behaviors that require understanding a system at multiple levels, with multiple causes interacting to create complex emergent effects (corresponding to level 5 in Brown, 2005)

Source: From Mislevey et al. (2014). Used with permission from the Institute of Play.

Figure 3 shows a couple of seconds worth of data, from one student in one challenge. What does one do with data like this? How does one make sense of the evidence when different students can follow different paths and use different strategies?

2. A Complex Adaptive Sociocognitive Systems Perspective

Dennett (1969) uses the term “person-level experience” for situations and events as people experience them and think about them, as they interact with the physical and social worlds; having a conversation, giving a presentation, or taking a test, as examples. These activities unfold over time. They are depicted in the middle layer of Fig. 4.



FIGURE 1.

Initial view of Jackson Citycaption. *Source:* From Mislevy et al. (2014). Used with permission from the Institute of Play.

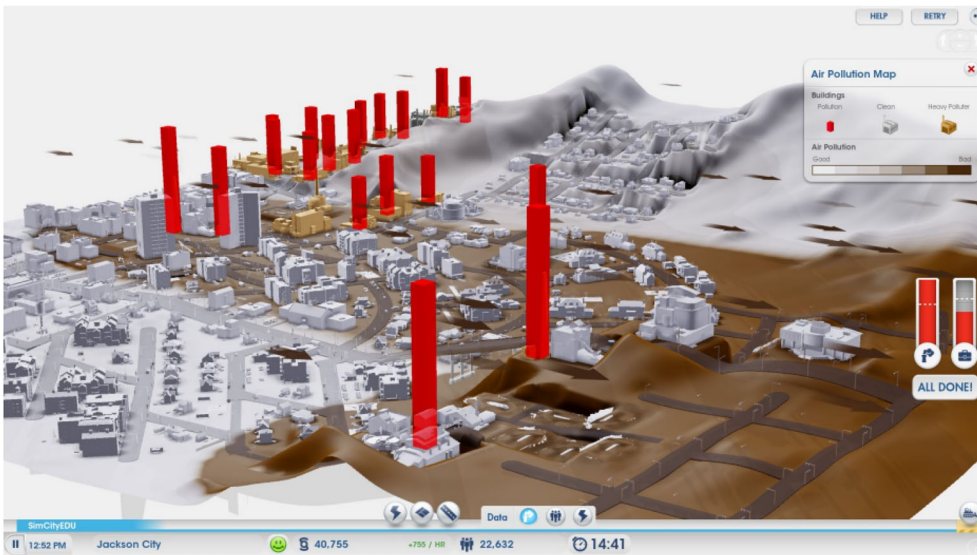


FIGURE 2.

Use of a tool to monitor pollution production. *Source:* From Mislevy et al. (2014). Used with permission from the Institute of Play.

For these interactions to be meaningful, much more must be going on at two levels, across people and within people. First, an interaction only makes sense by building on regularities across many such interactions, each unique, now and in the past—regularities that have to do with language, culture, and substance of an interaction, or LCS patterns for short. They are depicted in the top layer of Fig. 4. Every situation builds around many LCS patterns, of different kinds, at different grain sizes. Some of them a person is consciously aware of and thinks and acts through. Many more of them people are not aware of but think and act through them nevertheless. These

```

<stateInfo>
  <Init Fragment="600" Style="blue" Language="ENG" />
  <FSMStates Count="7">
    <finishButtonFSM>Stop Flashing</finishButtonFSM>
    <popupSubmitButtonFSM>Stop Flashing</popupSubmitButtonFSM>
    <submitButtonFSM>Stop Flashing</submitButtonFSM>
    <continueButtonFSM>Stop Flashing</continueButtonFSM>
    <nextButtonFSM>Stop Flashing</nextButtonFSM>
    <timeoutFSM>Clear</timeoutFSM>
    <playmakerFSM>Finish</playmakerFSM>
  </FSMStates>
  <OtherVars Count="3">
    <F6_Response>A</F6_Response>
    <F6_Reason1>asdf</F6_Reason1>
    <F6_Reason2>asdf</F6_Reason2>
  </OtherVars>
</stateInfo>
<ItemResult accessionNumber="TestAccNum" itemType="SBT" childItemAccessionNumber="176"
blockCode="TestBlockCode">
  <responseVariable cardinality="single" baseType="string">
    <candidateResponse>
      <value><![CDATA[["Selection of relevant questions":"Y,Y,Y,N"]]></value>
    </candidateResponse>
  </responseVariable>
  <responseVariable cardinality="single" baseType="string">
    <candidateResponse>
      <value><![CDATA[["How far away is the well?":"(a)Yes","Follow-up":"(a)There is probably not
enough water underground"]]></value>
    </candidateResponse>
  </responseVariable>
  <responseVariable cardinality="single" baseType="string">
    <candidateResponse>
      <value><![CDATA[["Wells in other villages?":"(b)No","Follow-up":null]]></value>
    </candidateResponse>
  </responseVariable>
  <responseVariable cardinality="single" baseType="string">
    <candidateResponse>

```

FIGURE 3.

Log file data from Jackson City activity. *Source:* From Mislevy et al. (2014). Used with permission from the Institute of Play.

regularities are culturally and historically contingent, meaning that LCS patterns can and do arise, evolve, and fade, and they vary over time and place, and across cultures and kinds of activities.

The existence of such regularities across people still isn't enough. Individuals must be able to recognize LCS patterns implicit in a situation, blend them with the particulars of that situation, and have some options for what to do next. They must have developed relevant *cognitive resources* through their personal history of experience—traces and generalizations from those specific situations, which were built around their own particular mixes of LCS patterns (Hammer et al., 2005; Young, 2009). These resources take the form of patterns of associations in the neural network of an individual's brain. They are depicted in the bottom layer of Fig. 4. While they are unique to a person, there can be similarities across peoples' experiences with respect to LCS patterns, and therefore in the cognitive resources they develop through their own experiences, and enable them to interact meaningfully.

There are three things to note about Fig. 4: (1) This is a complex adaptive system (Holland, 2006), so concepts from that field prove useful to understand interacting social phenomena (Byrne, Byrne 2002). (2) LCS patterns and individuals' resources are related through person-level activities and institutions, but they're different kinds of things. LCS patterns are emergent regularities in ways of acting and thinking over individuals, across myriad activities and intersecting communities (Sperber, 1996), while cognitive resources are individuals' attunements to such patterns as they have encountered instances of them through their experiences. (3) There are no constructs or measurement model θ s, which are proficiency variables that are a hallmark of measurement models such as item response theory (IRT) and cognitively diagnostic models (Borsboom, 2008; Van der Linden, 2017).

Here are some key implications of this complex adaptive sociocognitive system for construct-ing assessments, assessment arguments, and psychometric and data analytic methods:

- Every person-level situation builds around LCS patterns of many kinds and levels, and this includes assessment situations.

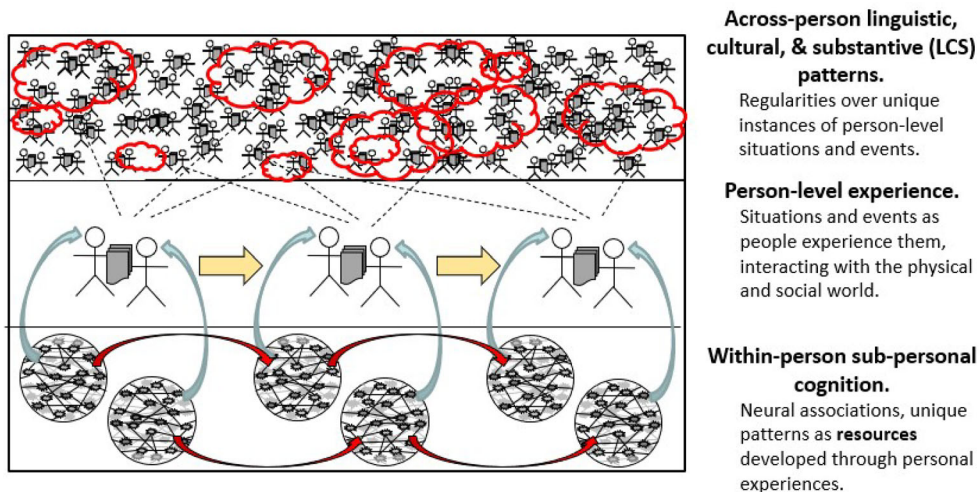


FIGURE 4.

A complex adaptive sociocognitive system. *Source:* Adapted from Mislevy (2012). Used with the permission of Educational Testing Service.

- An individual's experience of a situation assembles cognitive resources of many kinds, blended with features of that situation, much of which is below the level of consciousness (Kintsch, 1998).
- The cognitive resources each person develops are unique. They depend on personal history, in a person's milieu of experience.
- Regularities *across* persons can arise due to similarities that shape the situations they have experienced. Thus arise patterns in people's resources and actions. There are regularities and variation within and across people, and within and across situations.
- Such regularities and variation in a set of situations (e.g., tasks) as may arise—and as educators may arrange to arise—are the grist of constructs and of measurement modeling (see Gong et al., 2023, for an agent-based modeling illustration of the relation between sociocognitive processes and the parameters of an item response theory model).

3. Assessment Arguments

This section outlines the form of an assessment argument, discusses how it is fleshed out from a sociocognitive perspective, and extends the argument structure to interactive tasks.

3.1. The Basic Structure of Assessment Arguments

Figure 5 depicts the basic structure for an assessment argument. It captures the relationships expressed in Messick's quote, using concepts and representations developed by Wigmore (1913) and Toulmin (1958), and modernized by contemporary evidence scholars such as Anderson et al. (2005) and Schum (2001). Cronbach (1988), Messick (1989), Bachman and Palmer (2010), Kane (1992), Shepard (1993), and others have gainfully viewed assessment in terms of argument as concerning validity, and Mislevy et al. (2003), National Research Council (2001), and Wiley (1991) and others have done so concerning assessment design. Further extensions have addressed incorporation with learning models (Arieli Attali et al., 2019), assessment embedded in digital

games (Ke & Shute, 2015), and affordances provided from digital environments (e.g., Behrens et al. 2012).

In an assessment of any type, a user (perhaps a teacher, a student, or an admissions officer) desires to make a claim about a person (perhaps a student or themselves) in terms of some construal of capabilities of interest—constructs—based on some data. The same basic structure applies to assessments cast in any psychological perspective, including those under which educational measurement originated, namely trait, behavioral, and more recently, information processing. Greeno et al. (1997) argued that the sociocognitive perspective encompasses these other perspectives as special cases. I have discussed how, consequently, assessment arguments based on a sociocognitive perspective can similarly encompass arguments based on the other perspectives (Mislevy, 2018, Chapter 3–5).

When an analyst uses measurement models, the claim is represented in beliefs about the values of proficiency variables θ . (Yes, I did say there are no θ s in the actual complex system; I will return to this point presently.) The warrant is the set of beliefs and hypotheses that support this reasoning, such as generalizations, experience, scientific theories, and, in particular, psychometric measurement models. A central component of the main warrant in *SimCityEDU* is that a student able to reason at a given level in the progression in this context can generally reason at that level to tackle a particular challenge in which the underlying system and problem require it.

Alternative explanations are ways that despite the backing, a claim might not hold in a given case, even when backing supports the warrant as a generalization. As what Toulmin calls an informal argument rather than a logical argument, exception conditions can exist in an assessment argument. For example, in *SimCityEDU*, a student who does reason at Level 4 in some contexts might not do so in the Jackson City task because he is unfamiliar with the interface, misses some necessary information, or misunderstands the problem context.

Three kinds of data go into the main assessment argument, as shown in the three lower boxes. The first is aspects of a person's performance. This is not directly observed, but rather it is an evaluation of a unique performance, the cloud representing a person saying, doing, or making things in an assessment situation. The arrow from the cloud to the performance data indicates a sub-argument for the inference from the observation to this data, through the construal of the

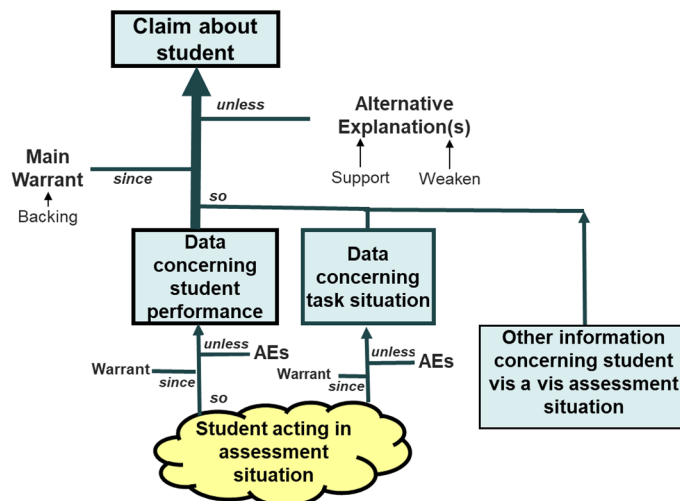


FIGURE 5.

The basic structure of an assessment argument. *Source:* From Mislevy (2012). Used with the permission of the Board of Regents of California.

construct. Alternative explanations may threaten that reasoning. I will return to how this sub-argument can become more complicated for the broader range of environments and interactions that assessment situations can comprise.

The next kind of data is features of the situation, because interpreting actions only makes sense in light of a situation—in evidence identification, as construed through the construct (which may not be the same as a student’s construal; in some cases, the intended identification is an inference about the student’s construal). It too is a sub-argument, with a warrant as to why the task situation satisfies the requirements of the main warrant to provide the desired evidence, and alternative explanations to investigate. In multiple-choice items, these features are built in by the test developer. In an interactive assessment, some features are built in, like the underlying system in a *SimCityEDU* challenge, but other features can be tailored to a student, and still other features of a situation a student is working in will emerge as the interaction between the system and the individual unfolds.

The third kind of data is additional knowledge about an individual in relation to the situation, because this knowledge conditions the inferences one can draw and the alternative explanations one must consider. Of all the many LCS patterns that are involved in the task for perception, action, and performance, any of them that are necessary but ancillary to the targeted construct can hamper some students and thus generate alternative explanations. The same performance in the same task holds different evidentiary value for you, for example, for inference about a student in your classroom when you know what they’ve been studying than it does for a user who doesn’t have this knowledge.

3.2. *Assessment Arguments from a Sociocognitive Perspective*

Assessment argumentation from a sociocognitive perspective has the same structure as under trait, behavioral, and information-processing perspectives that measurement modeling evolved under, but now every element is further informed by a sociocognitive perspective (Mislevy, 2018, Ch. 4 & 5). Assessment claims can still be organized around such constructs, but analysts and users are aware that constructs are elements of the model the analysts and users are employing, not an existing well-defined attribute possessed by a student. That is, they are external actors’ characterizations of students in terms of behavioral consistencies as the designers and users construe them. This construal, as well as evidence about such claims, is embedded in the designer’s sociocultural milieu, which need not correspond well to that of various students, in potential ways that bring about alternative explanations. Different actors who work from different psychological perspectives, have different purposes, or operate in different milieus may reason through different constructs.

Warrants framed in terms of constructs are to be understood in terms of resources and LCS patterns. When an argument is cast in terms of trait, behavioral, or information-processing perspectives, the warrant includes the presumption that such an approximation suits the contexts, populations, and purposes at issue. Backing includes theory, research, and experience that ground this component of the warrant for the application at issue.

Note that while such backing may support the warrant as generally applicable, it may not hold for given individuals, groups, or populations in the application at hand. Analysts and users become aware of the importance in the argument of the dependence of contexts and the interplay with students’ histories of learning and current performances. The analysts and users are thus alerted to alternative explanations that arise from atypical student backgrounds or LCS patterns necessary but ancillary task demands (Messick’s sources of construct irrelevant variance). For example, in measurement modeling the presence of differential item functioning (DIF) and person misfit suggest that an alternative explanation may be at play to cast doubt on the usual interpretation of scores.

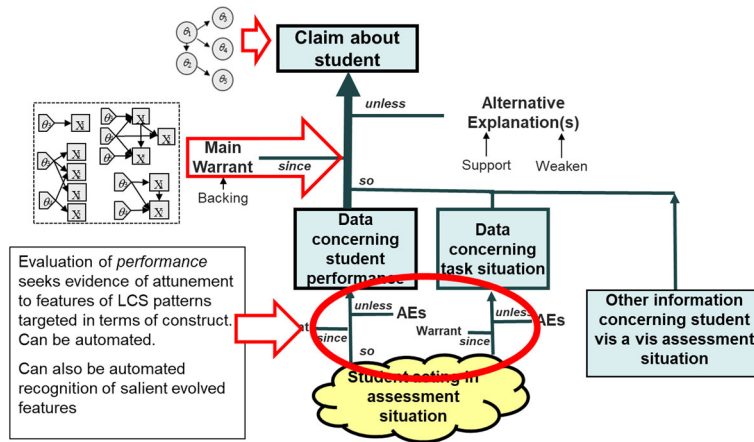


FIGURE 6.

Locations of psychometric models and evidence-identification methods. *Source:* Adapted from Mislevy (2012). Used with the permission of the Board of Regents of California.

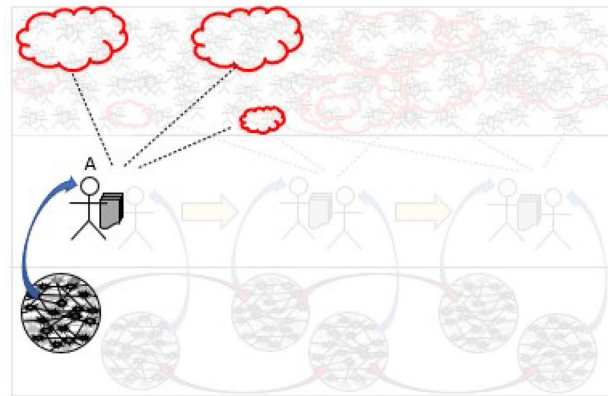
Evaluation of *performances* seeks evidence of students' attunement to features of targeted LCS patterns and the activities they draw on, as suggested in their actions. The warrant in this sub-argument says why the evaluation procedure generally provides data for the claim about the construct; alternative explanations range from incorrect answer keys in multiple-choice items, to aberrant human ratings, to missing unusual but insightful problem-solving sequences in a problem-solving situation.

An interpretation of a performance depends on an evaluation (perhaps implicit) of *features of the task situation*, because moment-to-moment actions in situations are evaluated in light of targeted practices and LCS patterns. This can require a much finer grain size for interactive and collaborative assessments than familiar ones. This is a grain size that cognitive modeling and situative psychology address naturally and are employed explicitly or implicitly in automated evaluation procedures (Yan et al., 2020).

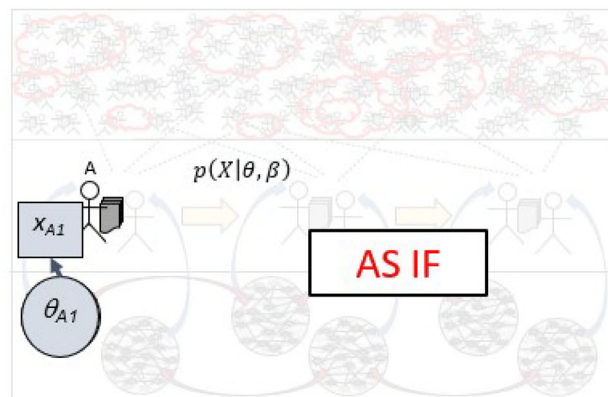
Figure 6 looks ahead to the principal places where psychometric measurement models fit into an argument when instantiated as an operational assessment. When claims are framed in terms of constructs and approximated in terms of values of θ s in a measurement model, the conditional probability distributions from hypothesized θ s to observable X s are an essential part of the reasoning, hence the warrant. A user reasons *as if* students possessed attributes θ and those θ s caused performance (Mislevy, 2018a). Evaluations X of salient data features (indicated by the oval in the figure) are obtained by evaluation procedures such as human ratings, data mining, machine learning, feature detectors, natural language processing (NLP), etc. I will return to this topic presently. Reasoning then flows back up through the measurement model conditional distributions $p(x|\theta)$ via Bayes theorem to update belief about θ . The warrant is that this measurement model approximation is adequate for the purpose at hand. Corresponding alternative explanations that arise are that the model does not fit sufficiently well to do so for an individual, for certain groups, or perhaps for anyone.

3.3. Extending the Structure to Interactive Tasks

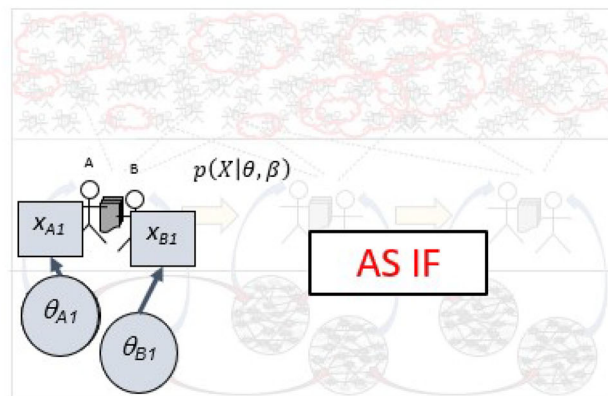
The basic structure discussed above suits a static task with a simple response, but not one with interaction and change during the performance, like a *SimCityEDU* challenge. This state



(a) Test-taker acting in task situation



(b) Test-taker action approximated by measurement model



(c) Test-takers' action approximated by measurement model

FIGURE 8.
Approximating relations between capabilities and actions with a measurement model.

The (Von Davier et al., 2021) *Computational Psychometrics* book discusses some more recent models.

I argued in *Sociocognitive Foundations of Educational Measurement* that from the perspective of model-based reasoning (Giere, 2004), between-persons educational measurement models are mathematical structures for expressing regularities and variabilities among persons, actions, and situations, shaped by theory and experience, built to serve understandings and purposes in contexts. From the perspective of probability-based inference, between-persons educational measurement models are Bayesian exchangeability structures (De Finetti, 1974) for managing evidence and inference in assessment. From a philosophical perspective, the θ s take constructive-realist (Messick, 1989) interpretations as regularities associated with persons in the instantiated model over some populations and both assessment and non-assessment situations. It is not that constructs, and by extension θ s, are real *per se*, but they do describe aspects of an ensemble model for patterns in actions, which are real, that arise from LCS patterns and practices in some milieu, which are also real, and within-person cognitive patterns, which too are also real, for functioning in that milieu. From a sociocognitive view, the sociohistorical locality and ongoing evolution of LCS patterns regarding education in culture will constrain the range and extent to which the forms and parameters of a model are suitable.

As a simple example to illustrate ideas, the Rasch IRT model for right/wrong test items addresses the observation that some people may tend to do better than others and some items are harder than others. The point here is just to relate measurement models in general to the argument structure and complex adaptive system diagram. Both the latent variable component θ and the data variable X of this model are quite simple: θ is a single continuous real-valued variable with higher values giving higher probabilities of a correct response, and X for any given item is a 0/1 response, however determined. This framing is shown in Panel b. Its parameters associated with people— θ s—and parameters associated with items— β s—give a first approximation to the details of a matrix of 1's and 0's. Salient features X of the performance are identified and characterized in ways I discuss more generally in a following section. The measurement model gives us probability distributions of possible values of X conditional on possible values of θ and β ; that is, $p(x | \theta, \beta)$. The θ s are the vehicle for model-based score interpretations and subsequent uses. A sociocognitive perspective cautions analysts that the overall patterns might differ with different collections of persons or situations. They are therefore on the lookout for systematic discrepancies for individuals or between groups that would suggest that alternative explanations hold for systematic patterns beyond those that a posited model can capture.

Panel c) adds another student taking the same assessment. Values of the same variable θ are used to characterize the capabilities of both students, and values of the same variable X are used to characterize the salient features of the responses of both students. Their cognitive resources may differ in a million ways and their performances may also differ in a million ways. But under the model, any differences in their performances are approximated as best as can be with only the possible values of X , and the differences in their capabilities as best as can be with only the possible values of θ .

The analyst then reasons provisionally *as if* this model were true, and draws inferences in terms of θ based on values of X . This is why the model is explicitly part of the warrant, and why it opens the door to alternative explanations. Of course the model is wrong, but the issue is whether it's good enough for the interpretations, the students, and the intended practical or scientific uses. The practical questions are then through which models, for what purposes, with which populations, facing what alternative explanations, is this approximation defensible? In a word, validation.

By designing, contextualizing, tailoring to circumstances and populations, and recognizing then averting, mitigating, or detecting when alternative explanations hold, assessors can in fact sometimes create an assessment system in which users can more or less reason as if the constructs

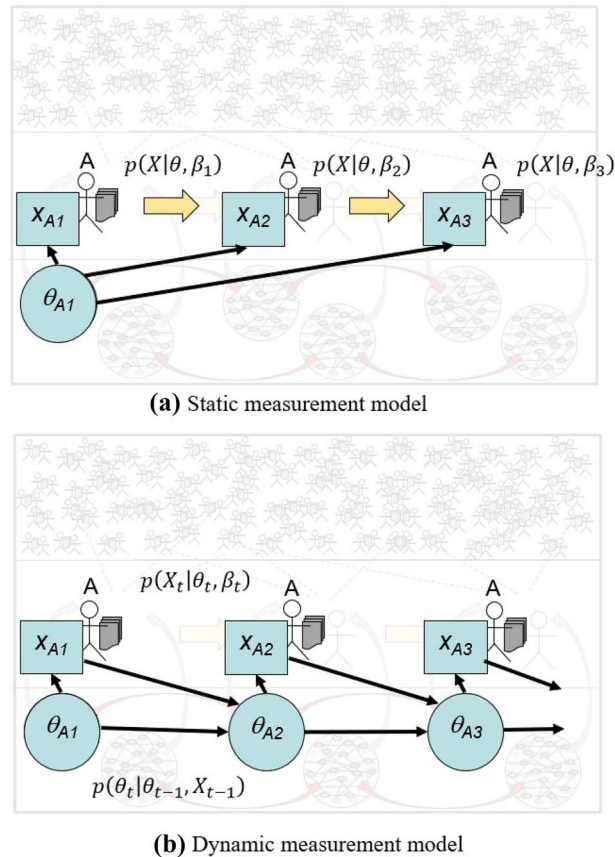


FIGURE 9.

Static and dynamic psychometric models approximating assessment performance in a complex adaptive sociocognitive system.

correspond directly to attributes of individuals and scores are measures of them. With familiar assessments, some of this pragmatic reasoning is built into good assessment practice, and some comes in by taking contexts, purposes, and populations into account in the task design, performance evaluation, and model construction. The more novel the assessment, the more useful an explicit framing like this becomes (Andrews-Todd et al., 2021), as with “stealth assessment” in game environments (Rahimi et al., 2023). It is equally useful when you are using an established assessment with a different interpretation, a new purpose, or a changing population (e.g., Fulcher & Davidson, 2009).

4.2. Modeling Multiple Observations

In large-scale testing, the usual assumption has been that the changes in a person’s capabilities of interest are negligible while they interact with the assessment. Panel a) of Fig. 9 shows how the IRT measurement model framing plays out. The same but unknown value of θ for a person is presumed at all time points. The response spaces and situation features can vary across observations/tasks but are again characterizable only within the predetermined definitions of the respective X and β variables.

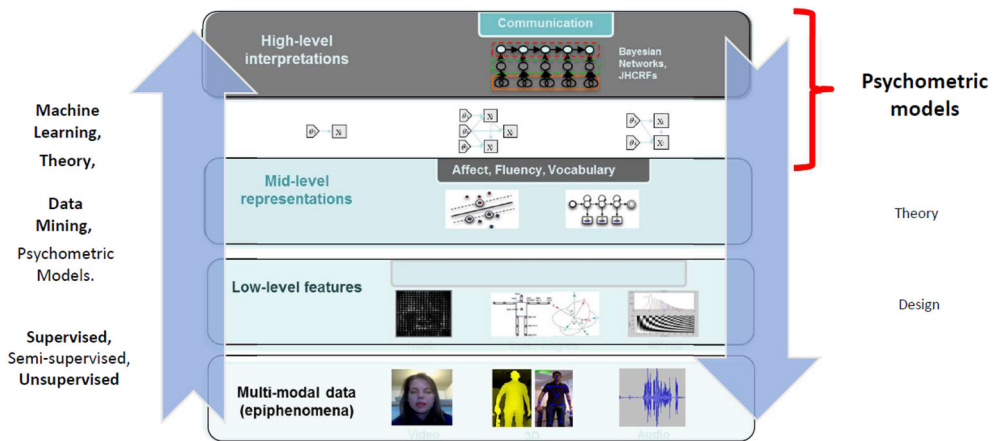


FIGURE 10.

Hierarchical evidence and task-feature evaluation. *Source:* Adapted from Khan (2017). Used with the permission of Educational Testing Service.

If it is relevant to model θ as changing through the experience (which is the whole point in learning systems) one needs a model that takes this into account. Mathematical psychologists developed dynamic models in the 1950s; later came production-rule learning models, and currently, there are models such as dynamic IRT (Glas & Verhelst, 1993), Bayesian model tracing (Desmarais & Baker, 2012), and partially observed Markov processes (Halpin et al., 2021). Panel b) of Fig. 9 shows a structure in which a person's θ at a given time point depends on both their previous θ and their performance at the previous time point.

5. Evidence Identification from a Sociocognitive Perspective

Now let's look more closely at the rationale and the processes of identifying and characterizing evidence from a more complicated, interactive performance in a possibly-evolving situation. Ultimately an analyst wants to interpret actions in terms of evidence for claims about constructs, which can be instantiated in terms of perhaps multivariate proficiency variables θ in a latent variable model. This section describes in general terms a hierarchical evidence-identification structure as it applies to a given performance. The following section shows how it applies in a *SimCityEDU* challenge like Jackson City.

What is initially captured in a complex performance as with a simulation- or game-based task is low-level data such as mouse clicks, drag-and-drop screen locations, and objects and connections in a construction. Modeling directly from the lowest-level observations to the highest-level constructs usually does not work well. For this reason, multiple layers of processing are typically employed in interactive assessments of any complexity. Figure 10 depicts a generic evidence-identification process, which allows for a sequence of successively refined processing stages: targeted latent variables at the top and lowest-level data at the bottom. This layering may be explicit, as with feature detectors at the lowest level providing input to a Bayesian inference network, or implicit, as with neural networks with multiple hidden layers (de Klerk et al., 2015; Yan et al., 2020).

Many evidence-identification techniques are being employed today in various assessment products and projects. At the left of the layers in Fig. 10 are some of the more bottom-up procedures that are used, including data mining and machine learning. The *Computational Psychometrics*

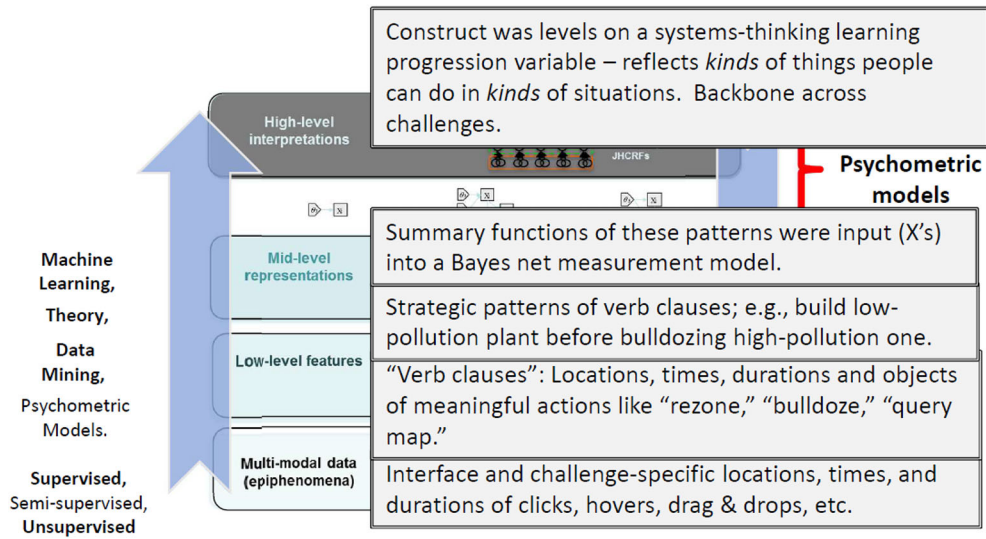


FIGURE 11.

Evidence identification processes in SimCityEDU. *Source:* Adapted from Khan (2017). Used with the permission of Educational Testing Service.

book and the *Handbook of Automated Scoring* mentioned earlier provide many details and examples. At the right at some of the more top-down ones, like psychometric models and theory-based designed-in situation features and affordances to bring about evidence-bearing opportunities. Note that LCS patterns—features of situations, meanings, and actions—are involved in designing the interface, underlying system, and affordances. LCS patterns are involved also in recognizing patterns of action at a finer grain size to parse and evaluate evidence of test-takers' capabilities.

In operation, low-level data is initially captured, and reasoning is successively upward. Moving up each layer is through reasoning that also can be analyzed in the same argument structure I have been using: What is the data coming in at that step, what is the intermediate claim coming out and being passed up to the next layer, what is the warrant for this stage and how well is it backed, and what alternative explanations is it vulnerable to? With inspectable models, one can examine the procedures and the reasoning using the structure of an assessment sub-argument with its substance as particularized LCS patterns.

6. Evidence Identification and Measurement Modeling in SimCityEDU

6.1. Evidence Identification

Figure 11 overlays the generic evidence-identification figure with the stages employed in the Jackson City challenge. The highest level is a summary of the salient aspects of the performance, as the X variables as evidence from this challenge about the targeted construct, operationalized as θ in an ordered latent class model defined by the systems-thinking learning progression (Table 1). That construct was levels on a systems-thinking learning progression variable, which reflects *kinds* of things people can do in *kinds* of situations. In a given performance on a given challenge, like Jackson City, the stages of the evidence-identification process produce an indication of the level exhibited in that performance.

At the lowest level of the figure is a raw data stream, consisting of interface-specific and sometimes current-situation-specific locations, times, and durations of clicks, hovers, drag &

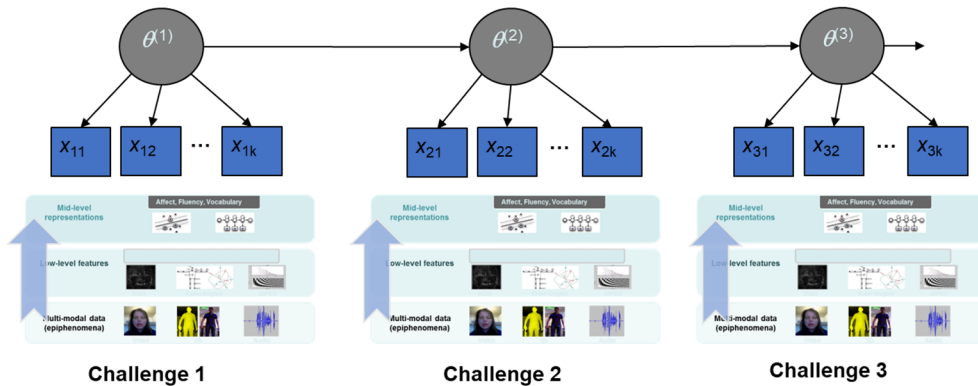


FIGURE 12.
Latent variable measurement model for SimCityEDU.

drops, etc. These are vital to the simulation’s calculations, but it is not the terms in which students think when they are playing. They are not even aware of this level of activity taking place “under the hood” of the game. Nor do they need to be.

The next level up is so-called verb clauses in the game: Locations, times, durations, and objects of actions that are meaningful in the game semantics, such as “bulldoze this building” or “query pollution map.” These are the terms students *do* think in. This is the level of the data snippet in Fig. 3.

The next level higher is strategic patterns of verb clauses; e.g., build a new low-pollution plant before you bulldoze an old high-pollution one. Identifying such patterns in evolved situations where they are appropriate is data, clues, about a student’s level of thinking about the city’s underlying pollution & jobs system. Some of these were hypothesized, then later fine-tuned with pilot data, using the theory of the proficiencies and the design of the game. Others were developed subsequently through data mining (DiCerbo et al., 2007).

Final challenge solutions and summary functions of strategic pattern usage in the challenge—counts, variety, and effectiveness as examples—were the vector of input variables X to the Bayes net psychometric model described next.

6.2. Measurement Modeling

The preceding section described hierarchical evidence-identification processes within each *SimCityEDU* challenge t , culminating in a vector of variables X_t that provide evidence about the level of thinking a student displayed in that performance. Figure 12 shows hierarchical evidence-identification chains for three successive challenges, for problems with increasingly complex underlying systems in the city.

In each problem, a student’s working through the task provides a vector of data summary variables X_t at Challenge t , to reflect a student’s level of system thinking during that challenge. Recall that the underlying system and problem to solve in a challenge were designed to require thinking at targeted levels as described in the learning progression (Table 1). The values of these X variables were modeled in terms of conditional probabilities given θ , as probabilities corresponding to expected performance by a student with proficiency at each given level in the learning progression. A student’s level in the progression variable θ was modeled as constant within a challenge but updated when the student moved on to the next challenge in accordance with probabilities at or above the level of the challenge crafted for a particular level in the learning progression.

As Fig. 10 suggests, there are a range of data analytic methods for identifying and evaluating evidence about a test-taker's capabilities from *within* a complex performance (Yan et al., 2020). Some involve probability-based modeling, others do not. There are, however, advantages to using probability-based latent variable models to synthesize such evidence *across* tasks, modes of performance, or segments within larger performances when test-takers experience different pathways or subtasks.

First, evidence about capabilities is synthesized in terms of posterior distributions for the proficiency variables θ . This approach subsumes traditional thinking about scores and measures but goes beyond it in ways that more complex performances and more ambitious inferences require (Mislevy & Gitomer, 1996; Rahimi et al., 2023). Further, validation approaches for examining the quality of “as if” reasoning about constructs that have developed over decades apply (Cronbach, 1971; Kane, 2006) and extend to new forms of data types and assessments (Ercikan et al., 2017; Zumbo et al., 2017). Probability modeling also affords real-time updating, quantification of evidence through posterior distributions for θ s, and, to investigate alternative explanations, model-critiquing methods with respect to individuals, groups, tasks, and background information.

7. Concluding Statement

This article sketches a view of educational measurement that adapts concepts and tools that originated under trait, behavioral, and information-processing perspectives on assessment, but as reconceived and extended along lines prompted by the more encompassing sociocognitive perspective. The complementary structuring draws on tools and concepts from evidentiary argumentation. The pillars of the proposed approach are as follows:

- Whether explicit or implicit, the psychological/social underpinning and substance of an assessment are essential to interpreting and using measurement model elements.
- A sociocognitive complex adaptive systems perspective connects disciplines involved in learning and assessment. These include technology, analytics, learning science, domain-based research, automated scoring, and assessment design.
- Measurement modeling remains useful in designing, critiquing, and using educational assessments—for managing issues of evidence and inference—but its design and use acquire situated meaning in and through sociocultural milieus.
- Argumentation structuring provides a framework for integrating and for working through the practical issues of assessment design, critique, and use. This structuring incorporates the measurement modeling framework.

This necessarily brief article offers an initial view of a sociocognitive approach to educational measurement, and the pillars above are more in the character of assertions than conclusions from the preceding discussion. Fuller details appear in the *Sociocognitive Foundations of Educational Measurement* book mentioned in the introduction. But even that is more an explication than a dialog with alternative views on the nature of constructs and variables, the nature and role of probability models, the sociohistorical nature of what people learn in cultures, the nature and acquisition of their capabilities, the sociocognitive interplay between inter- and intra-individual phenomena in a society, and the relations among these. This work is needed.

As I write, however, I can say that this argument structuring and sociocognitive perspective offer insights into familiar assessment and measurement practices. They make explicit evidentiary reasoning principles that appear to underlie familiar practices that worked well in the environment in which they evolved. I and others are finding that these principles, understood beyond the

particular forms they took, can be extended to incorporate advances in technology, analytics, and the psychology of learning.

Acknowledgments

This article builds from two presentations by the author: (1) The Career Award for Lifetime Achievement address “Further remarks on evidence and inference in educational assessment” at the International Meeting of the Psychometric Society (IMPS 2022), Bologna, Italy, July 12, 2022, and (2) A keynote address at the online workshop “Beyond Results 2021: From log data to valid inferences - Theory-based construction of process indicators,” hosted by the International Association for the Evaluation of Educational Achievement (IEA), together with the Leibniz Institute for Research and Information in Education (DIPF) and the Centre for International Student Assessment (ZIB), September 30th and October 1st, 2021. I am grateful to the Editor and two anonymous reviewers, whose insightful comments improved the exposition.

Declarations

Conflict of interest The author has no conflicts of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Anderson, T., Schum, D., & Twining, W. (2005). *Analysis of evidence*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511610585>
- Andrews-Todd, J., Mislevy, R. J., LaMar, M., & de Klerk, S. (2021). Virtual performance-based assessments. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methods for a new generation of educational assessment* (pp. 45–60). Springer. <https://doi.org/10.1007/978-3-030-74394-9>
- Arieli Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and process within assessment design. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2019.00853/full>
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press. <https://doi.org/10.1093/elt/ccq081>
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). An evidence-centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–54). Information Age.
- Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives*, 6, 25–53. <https://doi.org/10.1080/15366360802035497>
- Brown, N. J. S. (2005). *The multidimensional measure of conceptual complexity* (Tech. Rep. No. 2005-04-01). University of California, BEAR Center. <https://bearcenter.berkeley.edu/sites/default/files/report%20-%20mmcc.pdf>
- Byrne, D. (2002). *Interpreting quantitative data*. Sage Publications. <https://doi.org/10.4135/9781849209311>
- Cheng, B. H., Ructtinger, L., Fujii, R., & Mislevy, R. (2010). Assessing systems thinking and complexity in science (Large-Scale Assessment Technical Report 7). SRI International. http://ecd.sri.com/downloads/ECD_TR7_Systems_Thinking_FL.pdf
- Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Model and User-Adapted Interaction*, 12, 371–417. <https://doi.org/10.1023/A:1021258506583>

- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education. <https://doi.org/10.1016/j.compedu.2014.12.020>
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Erlbaum. <https://doi.org/10.1037/14047-004>
- Von Davier, A., Mislevy, R. J., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment*. Springer. <https://doi.org/10.1007/978-3-030-74394-9>
- De Finetti, B. (1974). *Theory of probability* (Vol. 1). Wiley. <https://doi.org/10.1002/9781119286387>
- de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, 85, 23–34. <https://doi.org/10.1016/j.compedu.2014.12.020>
- De Klerk, S., Veldkamp, B., & Eggen, T. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, 85, 23–34. <https://doi.org/10.1016/j.compedu.2014.12.020>
- Dennett, D. (1969). *Content and consciousness*. Routledge. <https://doi.org/10.4324/9780203092958>
- Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22, 9–38. <https://doi.org/10.1007/S11257-011-9106-8>
- DiCerbo, K., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, T. (2015). The role of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 319–342). Springer. https://doi.org/10.1007/978-3-319-05834-4_14
- Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning in the next generation of assessments*. Routledge. <https://doi.org/10.4324/9781315708591>
- Fisher, W. P. (2017). A practical approach to modeling complex adaptive flows in psychology and social science. *Procedia Computer Science*, 114, 165–174. <https://doi.org/10.1016/j.procs.2017.09.027>
- Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, 26, 123–144. <https://doi.org/10.1177/0265532208097339>
- Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71, 742–752. <https://doi.org/10.1086/425063>
- Glas, C. A. W., & Verhelst, N. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, 58, 395–415. <https://doi.org/10.1007/BF02294648>
- Gobert, J. D., Sao Pedro, M., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, 5, 153–185. <https://doi.org/10.5281/zenodo.3554645>
- Gong, T., Shuai, L., & Mislevy, R. J. (2023). Sociocognitive processes and item response models: A didactic example. *Journal of Educational Measurement*. <https://doi.org/10.1111/jedm.12376>
- Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–47). Simon & Schuster Macmillan. <https://doi.org/10.4324/9780203053874-8>
- Halpin, P., Ou, L., & LaMar, M. (2021). Time series and stochastic processes. In A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment* (pp. 209–230). Springer. https://doi.org/10.1007/978-3-030-74394-9_12
- Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89–120). Information Age Publishing. <https://doi.org/10.1201/9781410615749>
- Holland, J. H. (2006). Studying complex adaptive systems. *Journal of Systems Science and Complexity*, 19, 1–8. <https://doi.org/10.1007/s11424-006-0001-z>
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Praeger. <https://doi.org/10.1007/978-3-319-56129-5>
- Ke, F., & Shute, V. (2015). Design of game-based stealth assessment and learning support. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 301–318). Springer. <https://doi.org/10.1007/978-3-319-05834-4>
- Khan, S. M. (2017). Multimodal behavioral analytics in intelligent learning and assessment systems. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 173–184). Springer. <https://doi.org/10.1007/978-3-319-33261-1>
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511605968>
- Liu, L., Steinberg, J., Qureshi, F., Bejar, I., & Yan, F. (2016). Conversation-based assessments: An innovative approach to measure scientific reasoning. *Bulletin of the IEEE Technical Committee on Learning Technology*, 18(1), 10–13.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan. <https://doi.org/10.4324/9780203052341>
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. <https://doi.org/10.3102/0013189X0230002013>
- Mislevy, R. (2012). *Four metaphors we need to understand assessment*. Commissioned paper for The Gordon Commission on the Future of Assessment in Education. Educational Testing Service.

- Mislevy, R. J. (2018). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test data analytics and psychometrics: Informing assessment practices* (pp. 1–48). Information Age Publishing. <https://doi.org/10.1007/978-3-030-00033-0>
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. <https://doi.org/10.4324/9781315871691>
- Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., John, M., Bauer, M. I., Hoffman, E., von Davier, A. A., & Hao, J. (2014). *Psychometric considerations in game-based assessment*. Institute of Play.
- 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. <https://doi.org/10.1007/BF01126112>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–67. https://doi.org/10.1207/S15366359MEA0101_02
- Mislevy, R. J., Yan, D., Gobert, J., & Sao Pedro, M. (2020). Automated scoring with intelligent tutoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 403–422). CRC Press/Routledge. <https://doi.org/10.1201/9781351264808>
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. In J. Pellegrino, N. Chudowski, & R. Glaser (Eds.), Committee on the Foundations of Assessment. Board on Testing and Assessment, Center for Education Division of Behavioral and Social Sciences and Education. The National Academies Press.
- Paquette, L., Baker, R. S. J. D., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogof, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. Lecture notes in computer science. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems. ITS 2014* (Vol. 8474, pp. 1–10). Springer. https://doi.org/10.1007/978-3-319-07221-0_1
- Rahimi, S., Almond, R. G., Shute, V. J., & Sun, C. (2023). Getting the first and second decimals right: Psychometrics of stealth assessment. In M. P. McCreery & S. K. Krach (Eds.), *Games as stealth assessments* (pp. 125–153). IGI Global. <https://doi.org/10.4018/979-8-3693-0568-3>
- Schum, D. A. (2001). *The evidential foundations of probabilistic reasoning*. Northwestern University Press. <https://doi.org/10.1201/9781420035633>
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450. <https://doi.org/10.2307/1167339>
- Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Blackwell. <https://doi.org/10.1017/S0012217300007149>
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511840005>
- Van der Linden, W. J. (2017). *Handbook of item response theory* (Vol. 1–3). Chapman & Hall/ CRC Press. <https://doi.org/10.1007/978-1-4757-2691-6>
- Wigmore, J. H. (1913). *The principles of judicial proof: As given by logic, psychology, and general experience, and illustrated in judicial trials*. Little, Brown.
- Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75–107). Wiley. <https://doi.org/10.4324/9780203052341>
- Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press/Routledge. <https://doi.org/10.1201/9781351264808>
- Young, R. F. (2009). *Discursive practice in language learning and teaching*. Wiley-Blackwell. <https://doi.org/10.1017/S0272263109990453>
- Zumbo, B. D., & Huble, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research* (Vol. 26). Springer. <https://doi.org/10.1007/978-3-319-56129-5>

Manuscript Received: 6 DEC 2023

Accepted: 1 MAR 2024

Published Online Date: 3 APR 2024