Validity from the Perspective of Model-Based Reasoning[1]

Robert J. Mislevy
University of Maryland

December 9, 2008

[1] Presented at the conference "The Concept of Validity: Revisions, New Directions and Applications," University of Maryland, College Park, MD October 9-10, 2008. The work was supported by a grant from the Spencer Foundation.

To appear as

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications.* Charlotte, NC: Information Age Publishing.

Validity from the Perspective of Model-Based Reasoning

Abstract

From a contemporary perspective on cognition, the between-persons variables in trait-based arguments in educational assessment are absurd over-simplifications. Yet, for a wide range of applications, they work. Rather than seeing such variables as independently-existing characteristics of people, we can view them as summaries of patterns in situated behaviors that could be understood at the finer grainsize of sociocognitive analyses. When done well, inference through coarser educational and psychological measurement models suits decisions and actions routinely encountered in school and work, yet is consistent with what we are learning about how people learn, act, and interact. An essential element of test validity is whether, in a given application, using a given model provides a sound basis for organizing observations and guiding actions in the situations for which it is intended. This presentation discusses the use of educational measurement models such as those of item response theory and cognitive diagnosis from the perspective of model-based reasoning, with a focus on validity.

.

> A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure.
>
> Denny Borsboom, Gideon Mellenbergh, and Jaap van Heerden, 2004, p. 1.

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.
>
> Samuel Messick, 1989, p. 13.

> All models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> George Box and Norman Draper, 1987, p. 74.

Introduction

The concept of validity in educational assessment extends back more than a century (Sireci, 2008). The term was initially associated with the accuracy of predictions based on test scores. Concern with test content and with the meaning of scores gained attention in the middle of the century, with Cronbach and Meehl's (1955) "Construct validity in psychological tests" a watershed publication. More recent developments are the argument-based perspective noted in Messick's (1989) chapter in the third edition of *Educational Measurement* (Linn, 1989) and developed more fully by Kane (1992), and the use of cognitive theory to guide task design (Embretson, 1983). The present chapter contributes to these latter two lines of work, drawing on recent developments in cognitive psychology. In particular:

- A sociocognitive perspective on the nature of human knowledge provides insight into just what we are trying to assess.

- Research on the role of metaphors in cognition helps us understand the psychological, in conjunction with the formal, foundations of tools in the psychometric armamentarium.
- Studies of model-based reasoning in science provide a basis for understanding the activity of psychometric modeling.

Together, these lines of research are seen to support a constructivist-realist view of validity.

Preliminaries

Snow and Lohman's Assertion

In the 3rd edition of *Educational Measurement* (Linn, 1989), Messick (1989) defines a trait as "a relatively stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (p. 15). This is a common interpretation of the variables in the models of educational and psychological measurement. Snow and Lohman's chapter on cognitive psychology in the same volume proposes an alternative:

> Summary test scores, and factors based on them, have often been though of as "signs" indicating the presence of underlying, latent traits. … An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance.
>
> The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts.
>
> Snow & Lohman, 1989, p. 317.

This claim would seem to call into question the validity of inferences made through a conventional interpretation of test scores through educational and psychological measurement models.

Mixed-Number Subtraction

To illustrate ideas throughout the discussion, we will use an example drawn from the work of Kikumi Tatsuoka (e.g., Tatsuoka, 1983) on mixed number subtraction. Mixed-number subtraction problems require students to solve tasks such as $5\frac{1}{2}-3\frac{3}{4}$, $7\frac{2}{3}-\frac{1}{3}$, and $\frac{11}{8}-\frac{7}{8}$. A Rasch item response theory (IRT) model (Rasch, 1960/1980) often provides a reasonable fit to the right/wrong item responses of a group of middle-school students on a test of, say, twenty such tasks in open-ended format. The probability that Student *i* will respond correctly to Item *j*, or P_{ij}, is given as follows:

$$P_{ij} = P\left(X_{ij} = 1 \middle| \theta_i, \beta_j\right) = \frac{\exp\left(\theta_i - \beta_j\right)}{1+\exp\left(\theta_i - \beta_j\right)}, \quad (1)$$

where X_{ij} is the response of Student *i* to Item *j*, 1 if right and 0 if wrong; θ_i is a parameter for the proficiency of Student *i*; and β_j is a parameter for the difficulty of Item *j*. The less common multiplicative form of the model, an analogue of Newton's second law we will discuss in a later section, is for the odds of a correct response:

$$\frac{P_{ij}}{1-P_{ij}} = \xi_i / \delta_j, \quad (2)$$

where ξ_i=exp(θ_i) and δ_i=exp(β_i) from (1). IRT characterizations of students and items such as this are clearly simplifications, and they say nothing about the processes by which students answer items. They prove useful nevertheless for such purposes as tracking or comparing students' proficiency in this domain of tasks and quality-checking items.

In a series of publications in the 1980s, Tatsuoka and her colleagues developed a methodology for analyzing test item responses according to the rules – some correct, some incorrect – that students appeared to use to solve them (Birenbaum & K. Tatsuoka, 1983; Klein, Birenbaum, Standiford, & K. Tatsuoka, 1981; K. Tatsuoka, 1983, 1987,

1990; K. Tatsuoka & M. Tatsuoka, 1987). Extending earlier work by Brown & Burton (1978) to a statistical classification technique she called Rule Space, Tatsuoka characterized students in terms of the subset of rules that best seemed to explain their responses. Similarly, a binary skills latent class model (Haertel, 1989; Maris, 1999) provides an expression for the probability that Student *i* will answer Item *j* correctly, now in terms of which of *K* skills Item requires and which of these skills Student *i* can apply. Let $q_j = (q_{j1}, \ldots, q_{jK})$ be a vector of 0's and 1's for the skills Item *j* requires and $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})$ be a vector of 0's and 1's for the skills Student *i* can apply. Then the expression

$$P_{ij} = P\left(X_{ij} = 1 \middle| \eta_i, q_j\right) = \begin{cases} \pi_j & \text{if } \prod_k \eta_{ik}^{\,q_{jk}} = 1 \\ c_j & \text{if } \prod_k \eta_{ik}^{\,q_{jk}} = 0 \end{cases} \tag{3}$$

says that if Student *i* has all the skills Item *j* requires, the probability of getting it right is π_j, the true positive probability parameter for Item *j*, and if she lacks one or more of these skills, the probability is c_j, the false positive probability parameter for Item *j*. This is an example of what are now commonly called cognitively diagnostic models (CDMs; Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995).

While based on cognitive analyses of actual solutions, these models are also over-simplifications of students and solution processes. However, as Tatsuoka and her colleagues showed (also see VanLehn, 1990), they are useful for determining which concepts or procedures are useful for students to work on to improve their performance in the domain.

Questions

Snow and Lohman's assertion and the gainful use of different models for the same data raise philosophical questions about the nature of the parameters and probabilities in educational/psychological measurement models, the probabilities they entail, and of validity itself.

What is the nature of person parameters such as θ and η in latent variables models? Where do they reside?

What is the interpretation of the probabilities that arise from IRT and CDM models, and latent variable models in education and psychology more generally?

What are the implications of these observations for validity of models, assessments, and uses of them?

Some Relevant Results from Cognitive Science

Norman (1993) distinguishes between *experiential* and *reflective* cognition: "The experiential mode leads to a state in which we perceive and react to the events around us, efficiently and effortlessly. ... The reflective mode is that of comparison and contrast, of thought, of decision making. Both modes are essential to human performance" (Norman, 1993, pp. 15, 20). Both modes of cognition are involved in assessment. The first of three subsections that follow focuses on experiential cognition, from a sociocognitive perspective. It sheds light on the processes that Snow and Lohman suggested we consider to underlie test performances. The second discusses the roles of metaphor in cognition, and pertains to both experiential and reflective aspects. We see how metaphors ground the use of models in science, including in particular models such as those of IRT and CDM. The third describes model-based reasoning in greater detail, to set the stage for the discussion of IRT and CDMs from this perspective.

The Sociocognitive Perspective

Snow and Lohman's chapter is grounded in the cognitive revolution of the 1960s and 1970s), in which researchers such as Newell and Simon (1972) studied the nature of knowledge and how people might acquire, store, and retrieve it. The so-called first generation cognitive science drew on the metaphor of analytic computation, in the form of rules, production systems, task decompositions, and means–ends analyses. Contemporary work employs a connectionist metaphor to bring together results from psychology on learning, perception, and memory *within* individuals (e.g., Hawkins & Blakeslee, 2004) and fields such as linguistics and anthropology on the shared patterns of meaning and interaction *between* people (e.g., Gee, 1992; Strauss & Quinn, 1998). Linguist Dwight Atkinson (2002) calls this a sociocognitive perspective, to emphasize the interplay between the external patterns in the physical and the social world to which we

become attuned, and the patterns we develop and employ internally to understand and act accordingly.

One particular area in which these processes have been studied is reading comprehension. We can summarize the key ideas of Kintsch's (1998) construction-integration (CI) model for comprehension, and like Kintsch, take it as paradigmatic of comprehension more generally. Kintsch distinguish three levels involved in text comprehension, namely the *surface structure* of a text, the *text model*, and the *situation model*. The surface structure of a text concerns the specific words, sentences, paragraphs, and so on that constitute the text. The text model is the collection of interconnected propositions that the surface structures convey, and corresponds roughly to what might be called the literal meaning of a text. The situation model is a synthesized understanding that integrates the text model with the knowledge a reader brings to the encounter (also shaped by goals, affect, context, etc.), and constitutes that reader's comprehension of the text. Readers with different knowledge, affect, or purposes produce situation models that differ to varying degrees, and are unique due to each reader's history of experiences.

The construction (C) phase is initiated by features of stimuli in the environment and activates associations from long-term memory (LTM), whether they are relevant to the current circumstances or not. The associations include patterns of many kinds, from the forms of letters and grammatical constructions, to word meanings and discourse structures, to experiences with the subject matter at issue, such as the patterns and procedures in schemas for mixed number subtraction. The probability of activation of an element from LTM depends in large part on the strength of similarity of stimulus features and aspects of the elements of the schema. In the integration (I) phase, only the aspects of activated knowledge—both from contextual input and LTM—that are mutually associated are carried forward. The result, the *situation model*, is the reader's understanding of the text.

In assessment, the surface structure corresponds to the stimulus materials and conditions a task presents to the student. The text model is the intended meaning of that situation, within which the student is presumed to act. Situation models vary, often markedly, among students. A student may activate elements that are irrelevant from an

expert's point of view, and in unsystematic ways from one task to another, depending on idiosyncratic features of tasks and how they match up with the student's prior experiences (Redish, 2003). Kintsch and Greeno (1985), for example, studied how students solved, or failed to solve, arithmetic word problems using schemas from arithmetic, structures of the English language, and conventions for task design. This is the level of analysis that Snow and Lohman call attention to, and there are no θs or ηs in these processes within persons.

Despite the uniqueness of the processes within individuals, patterns of similarities do emerge. Individuals build up experiences that share similar features when they participate in instruction that uses common representational forms and terminology, when they work on similar problems using similar procedures, and when they talk with one another or read books based on the same concepts. As people acquire expertise in a domain, their knowledge becomes increasingly organized around key principles, and their perceptions and actions embody these shared ways of thinking. Although their experiences are unique, shared patterns in learning make for similarities in what students do in assessment situations. In the domain of mixed number subtraction, some students tend to solve more problems than others, and some items are harder than others as a result of the number and types of procedures they typically require. This observation motivates the idea of using of an IRT model to capture, express, and use these patterns for educative purposes. Patterns in what makes tasks hard and where students succeed and fail appear in relation to procedures and strategies. This motivates the use of a cognitive diagnosis model to guide instruction. Both models are wrong, to paraphrase the statistician George Box, but either might be useful in the right circumstances.

Metaphors in Human Cognition

Individual cognition is a unique blend of particular circumstances and more general patterns that are partly personal, due to our unique experiences, but partly shared with others, because they tap shared cultural models and because they build up as extensions of universal human experiences. With regard to the last point, one line of research in "embodied cognition" studies the roles of metaphor in cognition (Lakoff & Johnson, 1980, 1999). Lakoff asserts that our conceptual system, in terms of which we both think

and act, is fundamentally metaphorical in nature, building up from universal experiences such as putting things into containers and making objects move by bodily action. Our cognitive machinery builds from capabilities for interacting with the real physical and social world. We extend and creatively recombine basic patterns and relationships to think about everything from everyday things (a close examination of language shows it is rife with metaphor, much of which we do not even recognize as such) to extremely complicated and abstract social, technical, conceptual, and philosophical realms. The following sections consider four examples of metaphorical frames that are central to the use of models of educational and psychological measurement: containers, measurement, cause and effect, and probability. The section that follows this overview will show how these metaphors work together in measurement models in assessment.

Containers

The most fundamental metaphors are based on physical and spatial relationships in the world as humans, from birth, experience it. Examples are front and back ("We've fallen behind schedule"), moving along a path ("I'll start with a joke, move to my main points, and end with a moral"), up and down (with "up" as "good"), and the cause-and-effect metaphor discussed below. Containership is a basic physical and spatial relationship, where a container has an inside and an outside and is capable of holding something else. Dogs, apes, and parrots reason literally about containers, and employ them to achieve their ends. People reason metaphorically through the same structural relationships, continually and implicitly through the forms and the concepts based on containership relationships that are ubiquitous in all human languages, and formally and explicitly as the foundation of set theory and the classical definition of categories in philosophy (Lakoff & Johnson, 1999). As we noted earlier and will return to in a following section, latent class models build on the container metaphor.

Measurement

Measurement builds up from the physical experience of comparing objects in terms of their length or height. We experience "longer," "shorter," or "the same." Formalization from these simple foundations leads to more abstract concepts of catenation and

measuring devices for physical properties, then derived properties such as acceleration, axiomatization of measurement relationships, and the even more abstract relationships in the extension to conjoint measurement in social sciences (Michell, 1999). In abstract applications, the measurement metaphor posits variables that can be used to characterize all objects in a collection, each object is represented by a number, and the numbers can be used in further quantitative structures to characterize other events or relationships that involve the objects. The following paragraphs illustrate the role of measurement in quantitative structures within the cause-and-effect metaphor, specifically physical measurement in Newtonian mechanics and social-science measurement with the Rasch IRT model.

Cause and effect

Cause-and-effect reasoning is central to human reasoning in everyday life as well as in the disciplines. A dictionary definition is straightforward: One event, the cause, brings about another event, the effect, through some mechanism. Lakoff (1987, pp. 54ff) proposes that reasoning about causation extends from a direct-manipulation prototype that is basic to human experience; pushing a ball, for example, as shown in Figure 1a. He characterizes an idealized cognitive model for causation in terms of the following cluster of interactional properties:

1. *There is an agent that does something.*
2. *There is a patient that undergoes a change to a new state.*
3. *Properties 1 and 2 constitute a single event; they overlap in time and space; the agent comes in contact with the patient.*
4. *Part of what the agent does (either the motion or the exercise of will) precedes the change in the patient.*
5. *The agent is the energy source; the patient is the energy goal; there is a transfer of energy from the agent to patient.*
6. *There is a single definite agent and a single definite patient.*
7. *The agent is human.*
8. *a. The agent wills the action. b. The agent is in control of his action. c. The agent bears primary responsibility for both his actions and the change.*
9. *The agent uses his hands, body, or some instrument.*
10. *The agent is looking at the patient, the change in the patient is perceptible, and the agent perceives the change.'* (pp. 54-55).

[[Figure 1]]

Lakoff claims that the most representative examples of causation have all of these properties (e.g., Max broke the window). Less prototypical instances that we still consider causation lack some of the properties: indirect causation lacks Property 3, and billiard-ball interactions that characterize much reasoning in the physical sciences just have properties 1-6 (Figure 1b). Newtonian mechanics extends the cause-and-effect frame with sophisticated concepts such as mass, acceleration, and decomposition of forces, and adds a layer of quantitative relationships. Given a collection of springs and a collection of balls, for example, Newton's Second Law tells us how much acceleration results when each spring is used to propel each ball in terms of the spring's force, F_i and the ball's mass, M_j:

$$A_{ij} = F_i / M_j \,. \tag{3}$$

Latent variable models in educational and psychological measurement abstracting the cause-and-effect metaphor even further (Figure 1c). It is no coincidence that the multiplicative form of the Rasch model in Equation (2) mirrors Newton's Second Law in Equation (3). In his 1960 book *Probabilistic models for some intelligence and attainment tests,* Rasch explicitly lays out the analogy between (force, mass, acceleration) and (ability, item difficulty, probability of correct response). The measurement metaphor is intentional; how accurately and broadly it describes observations in given arenas of people and situations in a given application is to be determined. The latent class CDM model also draws on the causation metaphor, but with a different metaphor for the relationship between people and tasks, namely the container metaphor, and a correspondingly different quantitative layer.

Probability

Formal development of probability models began with systematic observations of games of chance. Shafer (1976) argues that these tangible, replicable situations ground reasoning about probability more generally. Kolmogorov's set theoretic basis of probability uses both the container metaphor and the measurement metaphor to describe what we see in repeated trials, and is an abstract foundation for a frequentist view of

probability. This interpretation of the metaphor considers probabilities to be a property of the world, induced by distributions of entities, mechanisms, or procedures. The same axioms ground reasoning using the same formal structure in further abstracted situations, as embodied in the personalistic or subjectivist Bayesian framework for probability (de Finetti, 1974; Savage, 1954). In this interpretation, probabilities are tools of the user, for reasoning about situations through a model, that is, an aspect of the formal, abstracted, specified, and situated applications of metaphors we discuss below as model-based reasoning. Either way, the use of the formal structures of probability models allow for reasoning about evidence and uncertainty in far more subtle and complex situations than unaided intuition can reckon with (Pearl, 1988; Schum, 1994).

The practical question in any application is whether the quantitative structure afforded by the probability framework, as particularized in terms of particular variables, models, and relationships, proves suitable for structuring reasoning in situations of interest. As we will see, the probability framework comes with some techniques that help one make this determination.

Model-Based Reasoning

A model is a simplified representation focused on certain aspects of a system (Ingham & Gilbert, 1991; cited in Gobert & Buckley, 2000). The entities, relationships, and processes of a model constitute its fundamental structure. They provide a framework for reasoning about patterns across any number of unique real-world situations, in each case abstracting salient aspects of those situations and going beyond them in terms of mechanisms, causal relationships, or implications at different scales or time points that are not apparent on the surface. To think about a particular situation for a particular purpose, scientists reason from principles in the domain to formulate a model that represents salient aspects of the situation, elaborate its implications, apprehend both anomalies and points of correspondence, and as necessary revise the model, the situation, or their theories in cycles of inquiry (Clement, 1989). Table 1 is based on Stewart and Hafner's (1994) and Gobert and Buckley's (2000) parsing of aspects of model-based reasoning.

[[Table 1: Aspects of model-based reasoning]]

Figure 2, based on Greeno (1983), suggests central properties of a model. The lower left plane shows phenomena in a particular real-world situation. A mapping is established between this situation and, in the center, structures expressed in terms of the entities, relationships, and properties of the model. Reasoning is carried out in these terms. This process constitutes an understanding of the situation, which can lead, through the machinery of the representation, to explanations, predictions, or plans for action. Above the plane of entities and relationships in the models are two symbol systems that further support reasoning in the model space, such as the matrix algebra and path diagram representations used in structural equation modeling. Note that they are connected to the real-world situation through the model.

[[Figure 2]]

The real-world situation is depicted in Figure 2 as fuzzy, whereas the model is crisp and well defined. This suggests that the correspondence between the real-world entities and the idealizations in the model are never exact. Not all aspects of the real-world situation are represented in the model. The model conveys concepts and relationships that the real-world situation does not. The reconceived situation shows a less-than-perfect match to the model, but is overlaid with a framework for reasoning that the situation itself does not possess in and of itself. These “surrogative inferences” (Swoyer, 1991) are precisely the cognitive value of a model (Suarez, 2004). A given model may, for example, support reasoning about missing data elements or future states of a situation.

It is particularly important that not everything in a real-world situation is represented in a model for that situation. Models address different aspects of phenomena, and can be cast at different levels. Different models address to different aspects of phenomena, and as such are tuned to reasoning about different problems. This observation underscores the user’s active role in model choice and construction, and the purpose for which the model is thought to be instrumental. One can examine aspects of transmission genetics with models at the level of species, individuals, cells, or molecules. One might use model water as molecular to study Brownian motion but as continuous to study flow through pipes (Giere, 2004). Newtonian mechanics has been superseded by relativity and

quantum theory, but it works fine for designing bridges. The constructivist-realist view holds that models are human constructions, but successful ones discern and embody patterns that characterize aspects of more complex real-world phenomena. Model-based reasoning is not just a dyadic relationship between a model and system, but a four-way relationship among a model, a situation, a user, and a purpose (Giere, 2004). In applied work, the issue is not a simple question of truth or falsity of a model, but of aptness for a given purpose.

The middle layer in Figure 2 is semantic, the narrative space of entities and relationships that are particularized to build stories to understand particular real-world situations. Metaphors play their roles here, as when we reason through the measurement metaphor when we use IRT and through the container metaphor when we use CDMs. In models that include quantitative layers, mathematical structures indicate forms of relationships, associates, and properties, and values of parameters in those models indicates the extent, strength, or variation within those forms as they might be used to approximate a given situation. These layers vary in the prominence across modeling enterprises and domains. Some models are strictly qualitative, and gain their power from the structures of entities, relationships, and processes they provide to reason through. Others, such as those in advanced physics, gain their power mainly through the mathematical relationships, and their users consider the narrative representations seriously inadequate on their own. Galileo famously said “Mathematics is the language with which God has written the universe.”

Models can additionally include probability components in two ways. The first is as substantive component of the model, when some of the relationships within a quantitative or qualitative layer are expressed in terms of probabilistic relationships. Item and person parameters in IRT models imply probabilities of responses, and variance components indicate among and ranges of data values or parameter values; these are inherently probabilistic relationships that obtain even if all data and parameters were known with certainty. The second is an overlay of the substantive model with a probabilistic layer that models the user’s knowledge and uncertainty about parameters and the structures within the substantive model, and the degree to which real-world observations accord

with the patterns the model can express. Modern psychometric models are probabilistic in both senses (Lewis, 1986).

The expression of a model's fit to data gives rise to an armamentarium of tools for exploring not only how well, but in what areas and in what ways the reconceived situation departs from the actual situation. A user can examine both the immediate model-data relationship and the quality of predications outside the immediate data such as predictions and appropriateness for new data. When the model is meant to serve a given purpose, it is of interest especially to see how well that purpose is subsequently served. This is the basis of predictive and consequential lines of validity argumentation in educational and psychological measurement.

Psychometrics as Model-Based Reasoning

A currently active and productive line of research in educational assessment is developing a view of assessment as evidentiary argument (e.g., Bachman, 2003; Kane, 1992, 2006; Mislevy, 2003, 2006). This work adapts tools and concepts from evidentiary reasoning (e.g., Schum, 1994; Toulmin, 1958; Wigmore, 1937) to help construct, critique, and validate assessments as instruments for reasoning from limited observations of students in particular situations to what they know or can do more broadly. The metaphorical and quantitative components of measurement models such as IRT and CDM serve as warrants in such applications.

An evidentiary argument is a series of logically connected claims or propositions that are supported by data by means of warrants (Toulmin, 1958). The claims in assessment arguments concern aspects of students' proficiency (Figure 3). Data consist of aspects of their performances in task situations and the salient features of those tasks. Warrants posit how responses in situations with the noted features depend on proficiency. Some conception of knowledge is the source of warrants, and shapes the nature of the claims a particular assessment is meant to support and the tasks and data needed to ground them (Mislevy, 2003, 2006). Alternative explanations weaken inference, and in arguments that rely on models this includes ways that the model ignores or misrepresents aspects of the situation that would in fact be relevant to the targeted inferences.

[[Figure 3—Assessment argument]]

An assessment based on the Rasch IRT model, for example, takes its IRT framework as its warrant. This includes both the metaphorical frame that characterizes persons and items by ability and difficulty parameters and the mathematical frame that gives probabilities of item response conditional on parameter values. Inferring the ability of a given student conditional on her observed responses (and good estimates of item parameters) requires reasoning back through the IRT model by means by Bayes theorem. An assessment based on the CDM takes the container metaphor as its narrative frame (you are in the container determined by your unobservable ηs) and conditional probabilities for item response given ηs; inferring a student's mastery of skills again required reasoning back through the model via Bayes theorem to obtain posterior probabilities for ηs, i.e., container membership.

Note again the metaphor drives not only the nature of the claim, but the aspects of the students' performances and the task situations that are deemed relevant. Figure 3 represents these determinations as embedded arguments, supported by warrants that justify the use of the IRT model in the context at hand. In mixed number subtraction, we might attend to different aspects of solutions in accordance with different psychological perspectives: correctness only from a trait perspective; specific answer, right or wrong, from an information-processing perspective in order to infer production rules; or adaptation of solution given hints when needed under a sociocultural perspective. Anticipation of what is important to observe similarly drives task construction, so the performance situation will be able to evoke the observations that are needed under the form of the warrant (Messick, 1994).

Alternative explanations in Toulmin's scheme condition an argument; they address ways that the data could be observed as was, yet the claim not be correct. Exception conditions to the warrant, for example, or misfit of the model in important respects. But how can one know what respects are important? This is where the purpose, or intended use, of the assessment comes into play. Bachman's (2003) calls the extension of the scheme to prediction, selection, instructional intervention, or program evaluation assessment the argument use argument (Figure 4). The claim emanating from the assessment argument is data for the assessment use argument. An inference about the student in the form of an IRT θ estimate or the most likely η vector in a CDM, is a

summary of selected aspects of performance as it can be expressed through the model—both as to its semantic content, in terms of the metaphorical frame, and its quantitative content, in terms of the mathematical structure associated with the model. This is the encoding of the information about the student as it will be employed in the use argument, which may, but need not, share the same view of the proficiency. Alternative explanations in the assessment use argument thus concern ways that the model-based inference from the assessment argument proper may be inadequate or misleading for the purpose at hand. In particular, neither the IRT nor the CDM model are faithful representations of the processes that produced responses, as viewed from the perspective of sociocognitive research.

[[Figure 4—Assessment use argument]]

Implications for Validity

We are now in a position to consider validity in educational assessment from the perspective of model-based reasoning. The discussion does so in relation to the quotations from Borsboom and Messick that appear at the beginning of this presentation.

Borsboom's Definition

Borsboom's definition of validity requires that the attribute an assessment is meant to measure *exists*, and that it *causes* variation in observations. We consider these points in turn.

Existence

As to existence, psychological research reinforces Snow and Lohman's doubts that trait interpretations are faithful representations of the cognition that produces test performances. Both research and experience in the learning sciences call attention to the situativity of cognition, as predictable from work like Kintsch's on reading comprehension. Examples include the context-dependent nature of reasoning in physics (e.g., Redish, 2003), language use (e.g., Chalhoub-Deville, 2003), and mathematics (Saxe, 1988). As a result, differences in contexts, formats, and degrees of familiarity affect the performance and consequent meaning of test scores both within and across

examinees. This phenomenon is well known to practitioners of educational and psychological measurement, in terms of method factors. What it suggests, however, is that attributes such as mixed-number subtraction ability are, as Snow and Lohman suggest, higher-level manifestations of more variegated processes rather than well-defined independently-existing properties of students.

Nevertheless, given specifications of conditions, task domains, and targeted testing populations, it may indeed be the case that students' propensities for actions in those tasks situations *can* be said to exist, and lead to exhibit patterns that can be approximated to some degree by models such as IRT and CDM. This is the constructivist realist position as it applies to practical educational and psychological measurement. Despite its inadequacy as a foundational explanation, the degree to which a given modeled representation suffices can be quite suitable for a given application. Snow and Lohman properly warn us against over-interpreting the model. An awareness of the finer-grained sociocognitive genesis of test performances continually suggests alternative explanations. Design principles and attention to context and use help us avoid inferential errors, and techniques such as model-fit analyses such as manifest and latent differential item functioning (Cohen & Bolt, 2005) and richer data such as talk-aloud solutions help us investigate them after the fact.

The first implication of a model-based reasoning perspective for Borsboom's definition of validity, then, is a softened view of an attribute's existence—one that brings in the four-way relationship among the model, the system, the user, and the purpose, and the adequacy of what the model does capture and the extent to which it does so, for the purpose at hand. Person parameters such as θ and η in psychometric models reside in the mind of the analyst rather than the mind of the subject, although their value depends on a resonance with discernable patterns in the real world associated with the subject.

Wiley (1995), like Borsboom, objected to Messick's inference-bound definition of validity. He argued that a test designer cannot be held responsible for validating all possible uses of a test, even though prior research and experience could ground the construction of a test that could be said to measure some attribute. I would hold that such circumstances do occur, but they hide assumptions about conditions, target populations, and purposes. As experience with certain kinds of tasks and accompanying models

accumulates and exhibits broader usefulness (they are said to be "fecund"), we gain confidence that the patterns they capture do reflect robustness that could be modeled in sociocognitive terms. However, I would hold also that we would be better served by viewing them as contingent outcomes of dissimilar processes at a finer grainsize, with uses therefore best confined to circumstances and purposes we should investigate and make ever more explicit. As discussed in the next section, the activities needed to do so comport nicely with the evolving tradition of test validation (Kane, 2006).

Causation

If θ and η in psychometric models reside in the mind of the analyst rather than the mind of the subject, how can they possibly be said to cause item responses *x*? The answer is straightforward from the perspective of model-based reasoning: when we have ascertained that the rationale and the evidence is sufficient to justify the use of an IRT or CDM model in a given application, the cause-and-effect metaphor is an appropriate structure to guide reasoning from observed responses to the targeted predictions, decisions, instructional feedback, and so on. This is so even though the IRT or CDM model is not a satisfactory explanation of responses, and even though the same structure applied to the same data could be misleading for different inferences.

Messick's Definition

Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 20). From the perspective of model-based reasoning, the attention to "adequacy and appropriateness of inferences and uses" is spot on. So is the criterion of "the degree to which empirical evidence and theoretical rationales" support this reasoning. I would disagree with Messick, however, and in so doing agree with Borsboom, Mellenbergh, and van Heerden (2004), that "Validity is not a judgment at all. It is the property being judged." From a model-based reasoning perspective, however, the property being judged must address the four-way relationship rather than the model-

system dyad. Either way, though, assessing it does indeed require an integrated evaluated judgment.[1]

And this is just what test validation has evolved to become in the special case of educational psychological measurement (Kane, 2006). Figure 5 shows that well-established lines of validation argumentation can be parsed in terms of model-based reasoning:

(1) Theory and experience supporting the narrative/scientific layer of the model. Although all models are wrong, we are more likely to have sound inference with models that are consistent with cognitive research and have proven practically useful in applications similar to the one at hand. As in physics, it is not that a model must be a faithful representation of a system, but that it capture the important patterns in ways that suit the intended inferences. Indeed, our uses of models like IRT and CDM are more likely to be successful if we *don't* believe the model is correct; we are more apt to be aware of alternative explanations, be diligent in model criticism, and stay within justifiable ranges of contexts, inferences, and testing populations.

(2) Theoretical and empirical grounding of task design. Not only is model formation a matter of construction and choice, so too is the design of the situations in which performance will be observed. What do the theory of knowledge and performance in a domain tell us about the features of tasks that are needed to prompt the targeted cognition? How do the features of the task situations align with features of future situations about which inference is intended? Embretson (1983) calls this the construct representation line of validity argumentation. Consistent with the spirit of model-based reasoning, design efforts can be based on finer-grained, higher-fidelity, or more encompassing psychological models than the simplified model that is used to synthesize data. Examples that use IRT or classical test theory include Embretson's (1998) cognitive-processing model for an analytic reasoning test, Bachman and Palmer's (1996) framework for tests that encompasses sociocultural

[1] Knowing the care with which Messick wrote, I have to believe he called validity the "judgment" rather than "the degree" intentionally. Was this a step from a constructivist-realist position toward a more radical constructivist position?

aspects of language use, and Katz's (1994) use of information-processing theory to ground task design in assessing proficiency in architectural design.

(3) Theoretical and empirical grounding of task-scoring procedures. We saw that an assessment argument also embeds an argument from students' performances to values of the observable variables that enter psychometric models. Another active line of research in educational and psychological measurement is increased attention to exactly what to identify and evaluate in performances, as motivated by more cognitively detailed understandings of the nature of proficiency and performance. The economically-motivated move to automated scoring of complex performances has spurred this development, because doing a practically better job has been found to require doing a scientifically better job (Williamson, Mislevy, & Bejar, 2006).

(4) Empirical evaluation of internal fit, predictions, and outcomes. A model's structure supports reasoning both about and beyond the data at hand. In models such as IRT and CDM with a quantitative, probabilistic layer as well as a semantic, metaphorical layer, we can ask how well the model's representation accords with the observed data. How good is the correspondence, where does it fail, and does it fail in ways that might be predicted by alternative explanations (e.g., differential item functioning)? These internal investigations must be supplemented by external investigations, such as correlations with other data and predictions of criterion performance that more directly address the quality of inferences obtained through the model.

[[Figure 5]]

Answers

We conclude with answers from the perspective of model-based reasoning to the questions posed earlier in the presentation about the nature of parameters, models, inferences, and validity.

What is the nature of person parameters such as θ and η in latent variables models? Where do they reside?

Person parameters in latent variables models are characterizations of patterns we observe in real-world situations (situations that we in part design for target uses), through the structure of a simplified model we are (provisionally) using to think about those situations and the use situations. They are in the heads of us, the users, but they aren't worth much unless they reflect patterns in examinees' propensities to action in the world. They are more likely to do so as (1) they accord with research and experience about the underlying nature of those propensities and actions, and (2) we design task situations and conditions of use in light of this emerging knowledge in such ways that will likely be robust with respect to the inevitable simplifications in the models. This view can be described as constructivist-realist.

What is the interpretation of the probabilities that arise from IRT and CDM models, and latent variable models in education and psychology more generally?

Probabilities are characterizations of patterns we observe in situations and our degree of knowledge about them, again through the structure of a simplified model we are (provisionally) using to think about those situations. The facts that (1) probabilities address the model space directly and only directly the real world though surrogative inference, and (2) different users and different purposes entail different models means that different probabilities can arise from different models. In addition to guiding inference through the model, the probabilistic layer of a quantitative model tools for seeing where the model may be misleading or inadequate.

What are the implications of these observations for validity of models, assessments, and uses of them?

A model-based reasoning perspective on the use of educational and psychological measurement is consistent with the currently dominant view of validity, which addresses "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). This is because model-based reasoning is concerned

with the four-way relationship among a model, a system, a user, and a purpose (Giere, 2004). Sources of validity evidence and lines of validity argumentation that have developed in the educational and psychological literature are nicely compatible with justifications of model-based reasoning in the scientific literature more generally.

References

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *The Modern Language Journal, 86,* 525-545.

Bachman, L.F. (2003). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1-34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Birenbaum, M., & Tatsuoka, K.K. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement, 20,* 17-26.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Box, G., & Draper, N. (1987). *Empirical model building and response surfaces.* New York: Wiley.

Brown, J. S., & Burton, R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Chalhoub-Deville, M. (2003) Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369-383.

Clement, J. (1989) Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning, and C. R. Reynolds (eds), *Handbook of Creativity: Assessment, Theory and Research* (pp. 341-381). New York: Plenum Press.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133-148.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity and psychological tests. *Psychological Bulletin, 52,* 281-302.

de Finetti, B. (1974). *Theory of probability (*Volume 1). London: Wiley.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Gee, J. P. (1992). *The social mind: Language, ideology, and social practice.* New York: Bergin & Garvey.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science, 71,* Supplement, S742-752.

Gobert, J. (2000). A typology of models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22,* 937-977.

Gobert, J. & Buckley, B. (2000). Special issue editorial: Introduction to model-based teaching and learning. *International Journal of Science Education, 22,* 891-894.

Greeno, J. G. (1983). Conceptual entities. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 227–252). Hillsdale, NJ: Lawrence Erlbaum Associates.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement, 26,* 301-321.

Hawkins, J, & Blakeslee, S. (2004). *On intelligence.* New York: Times Books.

Ingham, A. M. & Gilbert, J. K. (1991). The use of analogue models by students of chemistry at higher education level. *International Journal of Science Education, 13,*193-202.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 18-64). Westport, CT: Praeger.

Katz, I.R. (1994). Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram, N. Nersessian, & M. Recker (Eds.), *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society* (485-489). Mahwah, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W., & Greeno, J.G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109-129.

Klein, M.F., Birenbaum, M., Standiford, S.N., & Tatsuoka, K.K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions.* Research Report 81-6. Urbana, IL: Computer-based Education Research Laboratory, University of Illinois.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, G.P., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press

Lakoff, G., & Johnson, M. (1999) Philosophy in the flesh: The embodied mind and its challenge to western thought. New York: Basic Books.

Leighton, J.P. & Gierl, M. J. (Eds.) (2007). *Cognitive Diagnostic Assessment: Theories and Applications.* Cambridge: Cambridge University Press.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika, 51*, 11-22.

Linn, R.L. (Ed.) (1989), *Educational measurement* (3rd Ed.) New York: American Council on Education/Macmillan.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York: Cambridge University Press.

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2*, 237-258.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 257–305). Westport, CT: American Council on Education/Praeger Publishers.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Norman, D.A. (1993). *Things that make us smart.* Boston: Addison-Wesley.

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

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press (reprint).

Redish, E. F. (2003) *Teaching Physics with the Physics Suite*, John Wiley & Sons, Inc.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: John Wiley &. Sons, Inc.

Saxe, G.B. (1988). Candy selling and math learning. *Educational Researcher, 17*, 14-21.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: Princeton University Press.

Sireci, S. (2008). Packing and unpacking sources of validity evidence: History repeats itself again. Paper presented at the conference "The Concept of Validity: Revisions, New Directions and Applications," University of Maryland, College Park, MD October 9-10, 2008.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 263-331). New York: American Council on Education/Macmillan.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp 284-300). New York: Macmillan.

Strauss, C., & Quinn, N. (1998). *A cognitive theory of cultural meaning.* New York: Cambridge University Press.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*, Supplement, S767-779.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87,* 449-508.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K.K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*, 233-245.

Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K., & Tatsuoka, M.M. (1987). Bug distribution and statistical pattern classification. *Psychometrika, 52,* 193-206.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions.* Cambridge, MA: MIT. Press.

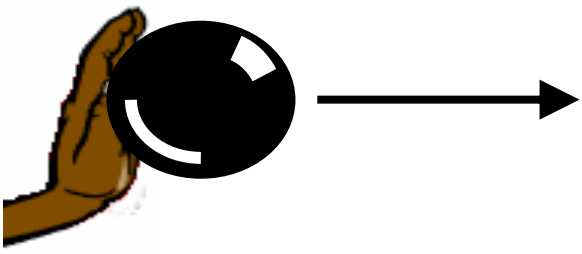
Wigmore, J.H. (1937). *The science of judicial proof (3rd Ed.).* Boston: Little, Brown, & Co.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75–107). Hillsdale, NJ: Lawrence Erlbaum.

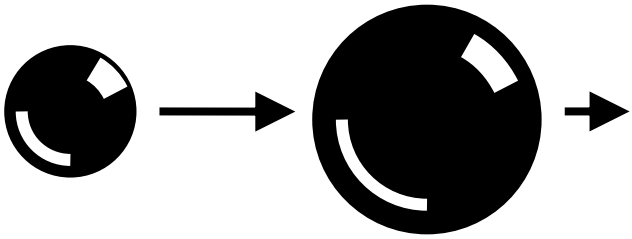
Williamson, D.M., Mislevy, R.J. & Bejar, I.I. (Eds.). (2006). *Automated Scoring of complex performances in computer based testing.* Mahwah, NJ: Erlbaum.

TABLE 1 Aspects of Model-Based Reasoning in Science

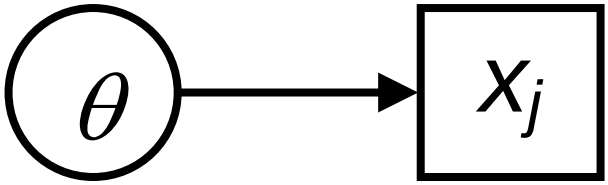
Model formation	Establishing a correspondence between some real-world phenomenon and a model, or abstracted structure, in terms of entities, relationships, processes, behaviors, etc. Includes scope and grain-size to model, and determining which aspects of the situation(s) to address and which to leave out.
Model elaboration	Combining, extending, adding detail to a model, establishing correspondences across overlapping models. Often done by assembling smaller models into larger assemblages, or fleshing out more general models with more detailed models.
Model use	Reasoning through the structure of a model to make explanations, predictions, conjectures, etc.
Model evaluation	Assessing the correspondence between the model components and their real-world counterparts, with emphasis on anomalies and important features not accounted for in the model.
Model revision	Modifying or elaborating a model for a phenomenon in order to establish a better correspondence. Often initiated by model evaluation procedures.
Model-based inquiry	Working interactively between phenomena and models, using all of the aspects above. Emphasis on monitoring and taking actions with regard to model-based inferences vis a vis real-world feedback.



a) Foundational experience



b) Newtonian mechanics



c) IRT model

Figure 1. Situations in which the cause and effect metaphor is employed

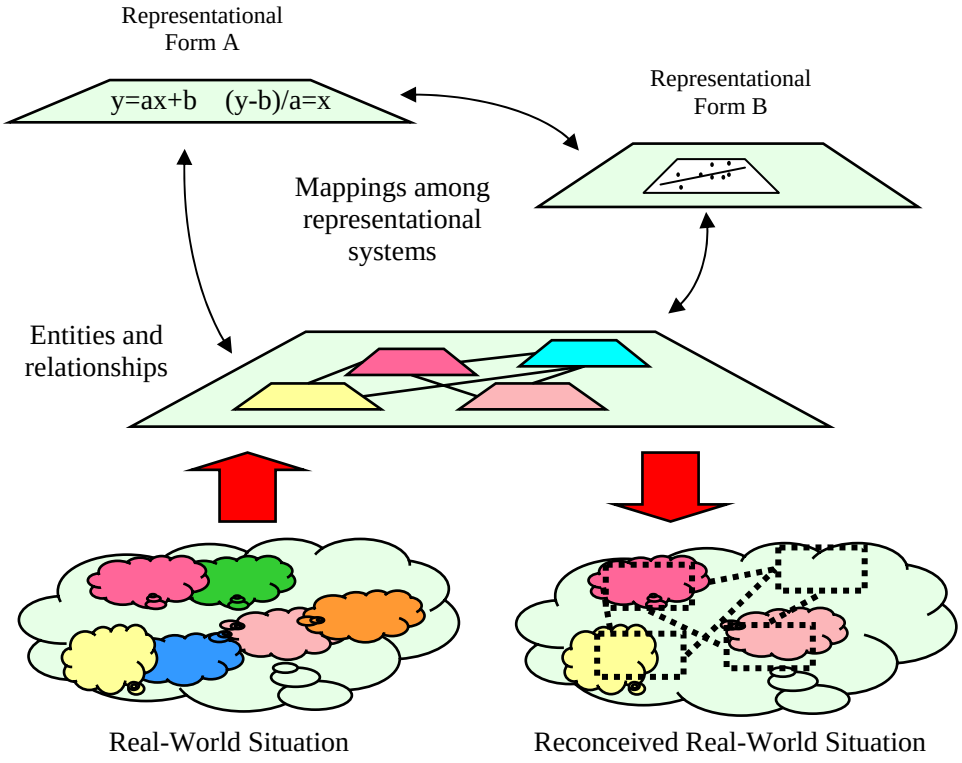


Figure 2. Reconceiving a real-world situation through a model

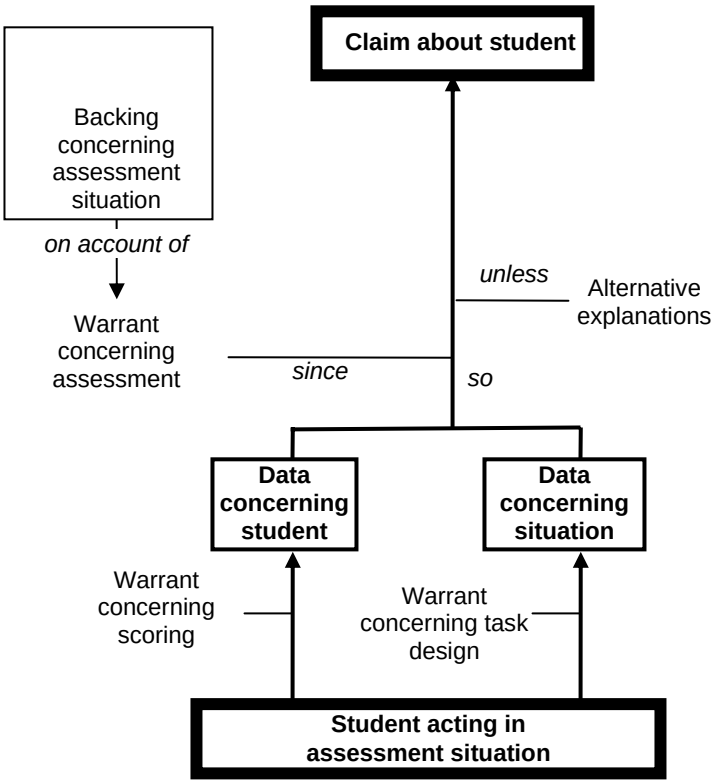


Figure 3. Assessment argument

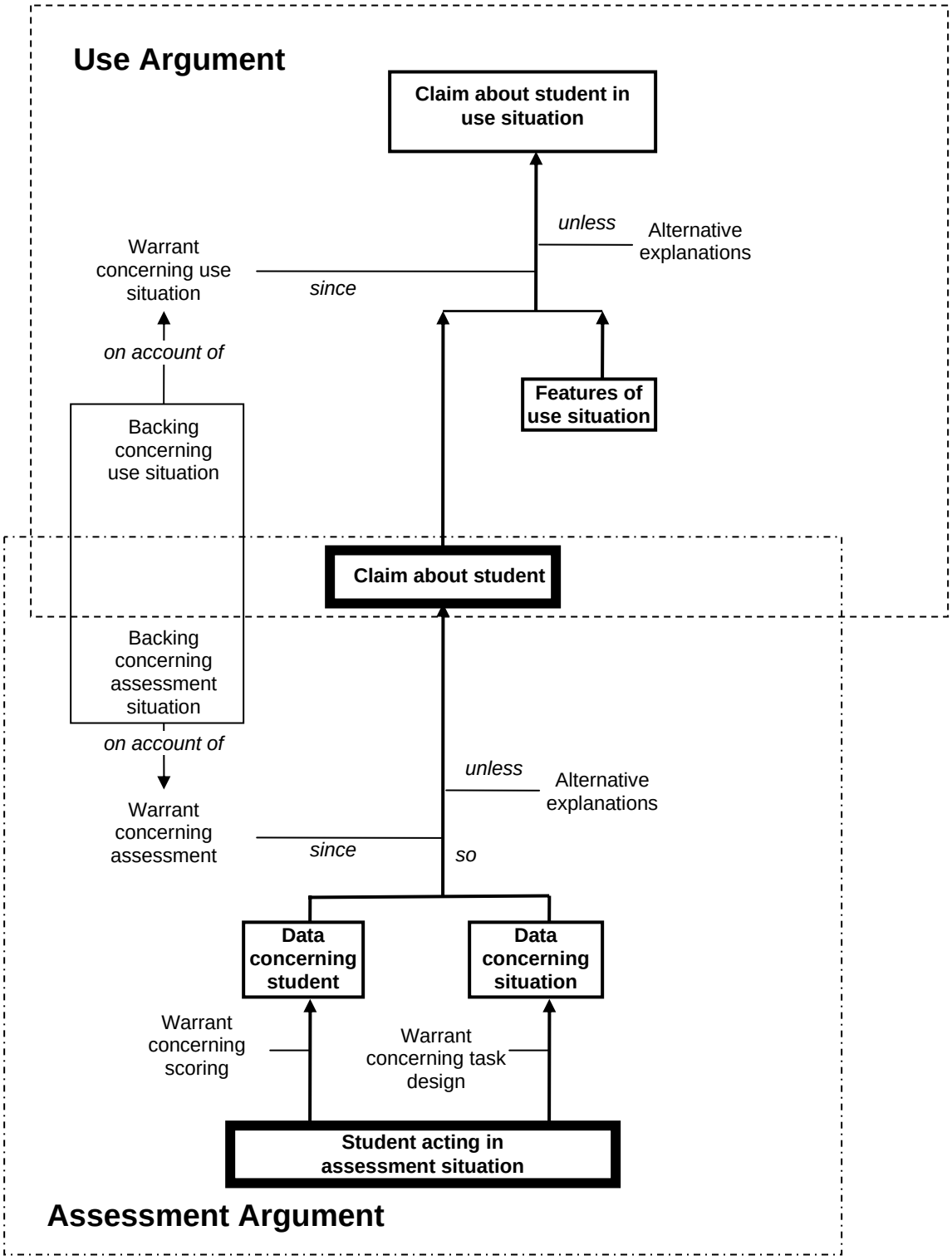


Figure 4. Assessment argument followed by assessment use argument

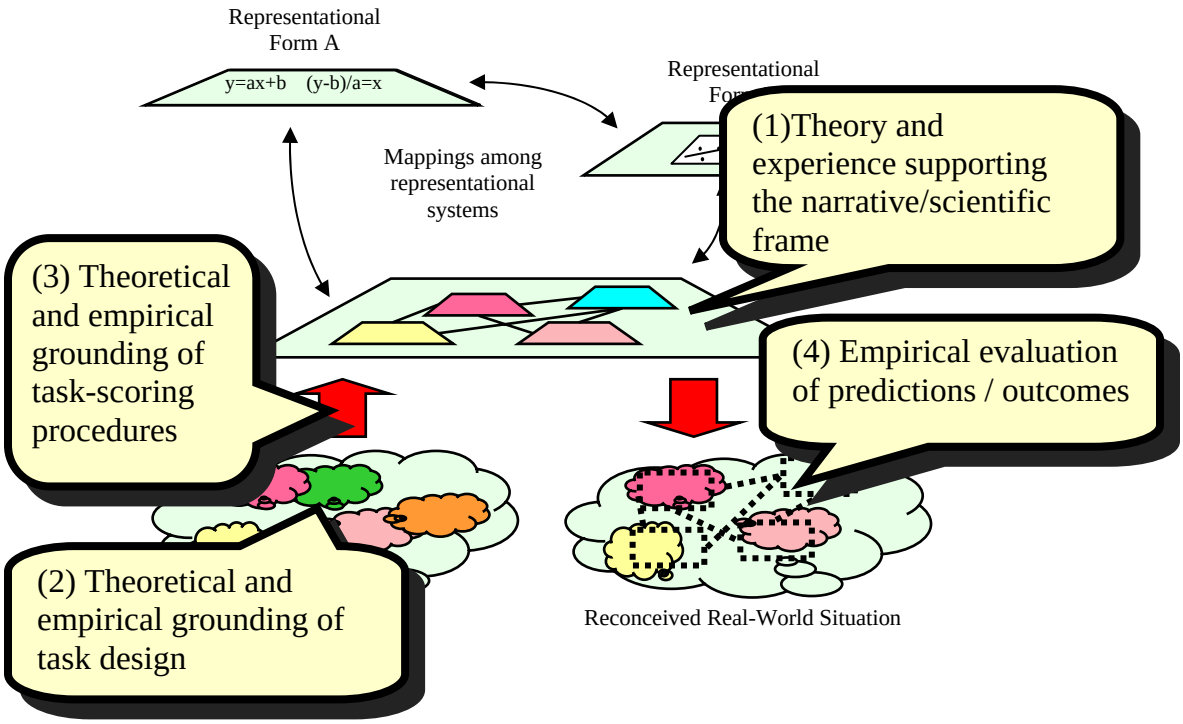


Figure 5. Lines of validity argumentation from the perspective of model-based reasoning