Psychometric Principles in Student Assessment

Robert J. Mislevy, University of Maryland

Mark R. Wilson, University of California at Berkeley

Kadriye Ercikan, University of British Columbia

Naomi Chudowsky, National Research Council

September, 2001

To appear in D. Stufflebeam & T. Kellaghan (Eds.), *International Handbook of Educational Evaluation.* Dordrecht, the Netherlands: Kluwer Academic Press.

ACKNOWLEDGEMENTS

This work draws in part on the authors' work on the National Research Council's Committee on the Foundations of Assessment. We are grateful for the suggestions of section editors George Madaus and Marguerite Clarke on an earlier version. The first author received support under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education. The second author received support from the National Science Foundation under grant No. ESI-9910154. The findings and opinions expressed in this report do not reflect the positions or policies of the National Research Council, the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, the National Science Foundation, or the U. S. Department of Education.

Psychometric Principles in Student Assessment

Abstract

In educational assessment, we observe what students say, do, or make in a few particular circumstances, and attempt to infer what they know, can do, or have accomplished more generally. Some links in the chain of inference depend on statistical models and probability-based reasoning, and it is with these links that terms such as validity, reliability, and comparability are typically associated—"psychometric principles," as it were. Familiar formulas and procedures from test theory provide working definitions and practical tools for addressing these more broadly applicable qualities of the chains of argument from observations to inferences about students, as they apply to familiar methods of gathering and using assessment data. This presentation has four objectives. It offers a framework for the evidentiary arguments that ground assessments, examines where psychometric principles fit in this framework, shows how familiar formulas apply these ideas to familiar forms of assessment, and looks ahead to extending the same principles to new kinds of assessments.

> [V]alidity, reliability, comparability, and fairness are not just measurement issues, but *social values* that have meaning and force outside of measurement wherever evaluative judgments and decisions are made.
>
> Messick, 1994, p. 2.

Overview

What are psychometric principles? Why are they important? How do we attain them? We address these questions from the perspective of assessment as evidentiary reasoning; that is, how we draw inferences about what students know, can do, or understand as more broadly construed, from the handful of particular things they say, do, or make in an assessment setting. Messick (1989), Kane (1992), and Cronbach and Meehl (1955) show the deep insights that can be gained from examining validity from such a perspective. We aim to extend the approach to additional psychometric principles and bring out connections with assessment design and probability-based reasoning.

Seen through this lens, validity, reliability, comparability, and fairness (as in the quote from Messick, above) are properties of an argument—not formulas, models, or statistics *per se*. We'll do two things, then, before we even introduce statistical models. We'll look more closely at the nature of evidentiary arguments in assessment, paying special attention to the role of standardization. And we'll describe a framework that structures the evidentiary argument in a given assessment, based on an evidence-centered design framework (Mislevy, Steinberg, & Almond, in press). In this way we may come to appreciate psychometric principles without tripping over psychometric details.

Of course in practice we do use models, formulas, and statistics to examine the degree to which an assessment argument possesses the salutary characteristics of validity, reliability, comparability, and fairness. So this presentation does have to consider how these principles are addressed when one uses particular measurement models to draw particular inferences, with particular data, for particular purposes. To this end we describe the role of probability-based reasoning in the evidentiary argument, using classical test theory to illustrate ideas. We then survey some widely-used psychometric models, such as item response theory and generalizability analysis, focusing how each is

used to address psychometric principles in different circumstances. We can't provide a guidebook for using all this machinery, but we will point out some useful references along the way for the reader who needs to do so.

This is a long road, and it may seem to wander at times. We'll start by looking at examples from an actual assessment, so the reader will have an idea of where we want to go, for thinking about assessment in general, and psychometric principles in particular.

An Introductory Example

The assessment design framework provides a way of thinking about psychometrics that relates what we observe to what we infer. The models of the evidence-centered design framework are illustrated in Figure 1. The *student model,* at the far left, concerns what we want to say about what a student knows or can do—aspects of their knowledge or skill. Following a tradition in psychometrics, we label this "θ" (theta). This label may stand for something rather simple, like a single category of knowledge such as vocabulary usage, or something much more complex, like a set of variables that concern which strategies a student can bring to bear on mixed-number subtraction problems and under what conditions she uses which ones. The *task model,* at the far right, concerns the situations we can set up in the world, in which we will observe the student say or do something that gives us clues about the knowledge or skill we've built into the student model. Between the student and task model are the scoring model and the measurement model, through which we reason from what we observe in performances to what we infer about a student.

Let's illustrate these models with a recent example—an assessment system built for a middle school science curriculum, "Issues, Evidence and You" (IEY; SEPUP, 1995). Figure 2 describes variables in the student model upon which both the IEY curriculum and its assessment system, called the *BEAR Assessment System* (Wilson & Sloane, 2000) are built. The student model consists of four variables at least one of which is the target of every instructional activity and assessment in the curriculum. The four variables are seen as four dimensions on which students will make progress during the curriculum. The dimensions are correlated (positively, we expect), because they all relate to "science", but are quite distinct educationally. The psychometric tradition would use a

diagram like Figure 3 to illustrate this situation. Each of the variables is represented as a circle—this is intended to indicate that they are unobservable or "latent" variables. They are connected by curving lines—this is intended to indicate that they are not necessarily causally related to one another (at least as far as we are modeling that relationship), but they are associated (usually we use a correlation coefficient to express that association).

==========
Insert Figures 1- 3 here
==========

The student model represents what we wish to measure in students. These are constructs—variables that are inherently unobservable, but which we propose as a useful way to organize our thinking about students. They describe aspects of their skill or knowledge for the purposes of, say, comparing programs, evaluating progress, or planning instruction. We use them to accumulate evidence from what we can actually observe students say and do.

Now look at the right hand side of Figure 1. This is the *task model.* This is how we describe the situations we construct in which students will actually perform. Particular situations are generically called "items" or "tasks".

In the case of IEY, the items are embedded in the instructional curriculum, so much so that the students would not necessarily know that they were being assessed unless the teacher tells them. An example task is shown in Figure 4. It was designed to prompt student responses that relate to the "Evidence and Tradeoffs" variable defined in Figure 2. Note that this variable is a somewhat unusual one in a science curriculum—the IEY developers think of it as representing the sorts of cognitive skills one would need to evaluate the importance of, say, an environmental impact statement—something that a citizen might need to do that is directly related to science's role in the world. An example of a student response to this task is shown in Figure 5.

How do we extract from this particular response some evidence about the unobservable student-model variable we have labeled Evidence and Tradeoffs? What we need is in the second model from the right in Figure 1—the *scoring model*. This is a procedure that allows one to focus on aspects of the student response and assign them to categories, in this case ordered categories that suggest higher levels of proficiency along

the underlying latent variable. A scoring model can take the form of what is called a "rubric" in the jargon of assessment, and in IEY does take that form (although it is called a "scoring guide"). The rubric for the Evidence and Tradeoffs variable is shown in Figure 6. It enables a teacher or a student to recognize and evaluate two distinct aspects of responses to the questions related to the Evidence and Tradeoffs variable. In addition to the rubric, scorers have exemplars of student work available to them, complete with adjudicated scores and explanation of the scores. They also have a method (called "assessment moderation") for training people to use the rubric. All these elements together constitute the scoring model. So, what we put in to the scoring model is a student's performance; what we get out is one or more scores for each task, and thus a set of scores for a set of tasks.

==========
Insert Figures 4-6 here
==========

What now remains? We need to connect the student model on the left hand side of Figure 1 with the scores that have come out of the scoring model—in what way, and with what value, should these nuggets of evidence affect our beliefs about the student's knowledge? For this we have another model, which we will call the *measurement model.* This single component is commonly known as a *psychometric model*. Now this is somewhat of a paradox, as we have just explained that the framework for psychometrics actually involves more than just this one model. The measurement model has indeed traditionally been the focus of psychometrics, but it is not sufficient to understand psychometric principles. The complete set of elements, the full evidentiary argument, must be addressed.

Figure 7 shows the relationships in the measurement model for the sample IEY task. Here the student model (first shown in Figure 3) has been augmented with a set of boxes. The boxes are intended to indicate that they are observable rather than latent, and these are in fact the scores from the scoring model for this task. They are connected to the Evidence and Tradeoffs student-model variable with straight lines, meant to indicate a causal (though probabilistic) relationship between the variable and the observed scores, and the causality is posited to run from the student model variables to the scores. Said

another way, what the student knows and can do, as represented by the variables of the student model, determines how likely it is that the students will make right answers rather than wrong ones, carry out sound inquiry rather than founder, and so on, in each particular task they encounter. In this example, both observable variables are posited to depend on the same aspect of knowledge, namely Evidence and Tradeoffs. A different task could have more or fewer observables, and each would depend on one or more student-model variables, all in accordance with what knowledge and skill the task is designed to evoke.

==========

Insert Figure 7 here

==========

It is important for us to say that the student model in this example (indeed in most psychometric applications) is not proposed as a realistic explanation of the thinking that takes place when a student works through a problem. It is a piece of machinery we use to accumulate information across tasks, in a language and at a level of detail we think suits the purpose of the assessment (for a more complete perspective on this see Pirolli & Wilson, 1998). Without question, it is selective and simplified. But it ought to be consistent with what we know about how students acquire and use knowledge, and it ought to be consistent with what we see students say and do. This is where psychometric principles come in.

What do psychometric principles mean in IEY? *Validity* concerns whether the tasks actually do give sound evidence about the knowledge and skills the student-model variables are supposed to measure, namely, the five IEY progress variables. Or are there plausible alternative explanations for good or poor performance? *Reliability* concerns how much we learn about the students, in terms of these variables, from the performances we observe. *Comparability* concerns whether what we say about students, based on estimates of their student model variables, has a consistent meaning even if students have taken different tasks, or been assessed at different times or under different conditions. *Fairness* asks whether we have been responsible in checking important facts about students and examining characteristics of task model variables that would invalidate the inferences that test scores would ordinarily suggest.

Psychometric Principles and Evidentiary Arguments

We have seen through a quick example how assessment can be viewed as evidentiary arguments, and that psychometric principles can be viewed as desirable properties of those arguments. Let's go back to the beginning and develop this line of reasoning more carefully.

Educational assessment as evidentiary argument[i]

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. Rarely do we have the luxury of reasoning with certainty; the information we work with is typically incomplete, inconclusive, amenable to more than one explanation. The very first question in an evidentiary problem is, "evidence about what?" *Data* become *evidence* in some analytic problem only when we have established their relevance to some conjecture we are considering. And this task of establishing the relevance of data and its weight as evidence depends on the chain of reasoning we construct from the evidence to those conjectures.

Both conjectures and an understanding of what constitutes evidence about them, arise from the concepts and relationships of the field under consideration. We'll use the term "substantive" to refer to these content- or theory-based aspects of reasoning within a domain, in contrast to structural aspects such as logical structures and statistical models. In medicine, for example, physicians frame diagnostic hypotheses in terms of what they know about the nature of diseases, and the signs and symptoms that result from various disease states. The data are patients' symptoms and physical test results, from which physicians reason back to likely disease states. In history, hypotheses concern what happened and why. Letters, documents, and artifacts are the historian's data, which she must fit into a larger picture of what is known and what is supposed.

Philosopher Stephen Toulmin (1958) provided terminology for talking about how we use substantive theories and accumulated experience (say, about algebra and how kids learn it) to reason from particular data (Joe's solutions) to a particular claim (what Joe understands about algebra). Figure 8 outlines the structure of a simple argument. The *claim* is a proposition we wish to support with *data*. The arrow represents inference, which is justified by a *warrant*, or a generalization that justifies the inference from the

particular data to the particular claim. Theory and experience provide *backing* for the warrant. In any particular case we reason back through the warrant, so we may need to qualify our conclusions because there are *alternative explanations* for the data.

[[Figure 8—basic Toulmin diagram]]

In practice, of course, an argument and its constituent claims, data, warrants, backing, and alternative explanations will be more complex than Figure 8. An argument usually consists of many propositions and data elements, involves chains of reasoning, and often contains dependencies among claims and various pieces of data. This is the case in assessment.

In educational assessments, the data are the particular things students say, do, or create in a handful of particular situations—written essays, correct and incorrect marks on answer sheets, presentations of projects, or explanations of their problem solutions. Usually our interest lies not so much in these particulars, but in the clues they hold about what students understand more generally. We can only connect the two through a chain of inferences. Some links depend on our beliefs about the nature of knowledge and learning. What is important for students to know, and how do they display that knowledge? Other links depend on things we know about students from other sources. Do they have enough experience with a computer to use it as a tool to solve an interactive physics problem, or will it be so unfamiliar as to hinder their work? Some links use probabilistic models to communicate uncertainty, because we can administer only a few tasks or because we use evaluations from raters who don't always agree. Details differ, but a chain of reasoning must underlie an assessment of any kind, from classroom quizzes and standardized achievement tests, to coached practice systems and computerized tutoring programs, to the informal conversations students have with teachers as they work through experiments.

The case for standardization

Evidence rarely comes without a price. An obvious factor in the total cost of an evidentiary argument is the expense of gathering the data, but figuring out what data to gather and how to make sense of it can also be dear. In legal cases, these latter tasks are usually carried out after the fact. Because each case is unique, at least parts of the

argument must be uniquely fashioned. Marshalling evidence and constructing arguments in the O.J. Simpson case took more than a year and cost millions of dollars, to prosecution and defense alike.

If we foresee that the same kinds of data will be required for similar purposes on many occasions, we can achieve efficiencies by developing standard procedures both for gathering the data and reasoning from it (Schum, 1994, p. 137). A well-designed protocol for gathering data addresses important issues in its interpretation, such as making sure the right kinds and right amounts of data are obtained, and heading off likely or pernicious alternative explanations. Following standard procedures for gathering biological materials from crime scenes, for example, helps investigators avoid contaminating a sample, and allows them to keep track of everything that happens to it from collection to testing. What's more, merely confirming that they've followed the protocols immediately communicates to others that these important issues have been recognized and dealt with responsibly.

A major way that large-scale assessment is made practicable in education is by thinking these issues through up front: laying out the argument for what data to gather and why, from each of the many students that will be assessed. The details of the data will vary from one student to another, and so will the claims. But the same kind of data will be gathered for each student, the same kind of claim will be made, and, most importantly, the same argument structure will be used in each instance. This strategy offers great efficiencies, but it admits the possibility of cases that do not accord with the common argument. Therefore, establishing the credentials of the argument in an assessment that is used with many students entails the two distinct responsibilities listed below. We shall see that investigating them and characterizing the degree to which they hold can be described in terms of psychometric principles.

- *Establishing the credentials of the evidence in the common argument.* This is where efficiency is gained. To the extent that the same argument structure holds for all the students it will be used with, the specialization to any particular student inherits the backing that has been marshaled for the general form. We will discuss below how the common argument is framed. Both rational analyses and large-scale statistical

analyses can be used to test its fidelity at this macro level. These tasks can be arduous, and they can never really be considered complete because we could always refine the argument or test additional alternative hypotheses (Messick, 1989). The point is, though, that this effort does not increase in proportion to the number of examinees who are assessed.

- *Detecting individuals for whom the common argument does not hold.* Inevitably, the theories, the generalizations, the empirical grounding for the common argument will not hold for some students. The usual data arrive, but the usual inference does not follow—even if the common argument does support validity and reliability in the main. These instances call for additional data or different arguments, often on a more expensive case-by-case basis. An assessment system that is both efficient and conscientious will minimize the frequency with which these situations occur, but routinely and inexpensively draw attention to them when they do.

It is worth emphasizing that the standardization we are discussing here concerns the structure of the argument, not necessarily the form of the data. Some may think that this form of standardization is only possible with so-called objective item forms such as multiple choice items. Few large-scale assessments are more open-ended than the Advanced Placement Studio Art portfolio assessment (Myford & Mislevy, 1995); students have an almost unfettered choice of media, themes, and styles. But the AP program provides a great deal of information about the qualities students need to display in their work, what they need to assemble as work products, and how raters will evaluate them. This structure allows for a common argument, heads off alternative explanations about unclear evaluation standards in the hundreds of AP Studio Art classrooms across the country, and, most happily, helps the students come to understand the nature of good work in the field (Wolf, Glenn, and Gardner, 1991).

Psychometric principles as properties of arguments

Seeing assessment as argument from limited evidence is a starting point for understanding psychometric principles.

Validity

Validity is paramount among psychometric principles, for validity speaks directly to the extent to which a claim about a student, based on assessment data from that student, is justified (Cronbach, 1989; Messick, 1989). Establishing validity entails making the warrant explicit, examining the network of beliefs and theories on which it relies, and testing its strength and credibility through various sources of backing. It requires determining conditions that weaken the warrant, exploring alternative explanations for good or poor performance, and feeding them back into the system to reduce inferential errors.

In the introductory example we saw that assessment is meant to get evidence about students' status with respect to a construct, some particular aspect(s) of knowledge, skill, or ability—in that case, the IEY variables. Cronbach and Meehl (1955) said "construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality is not operationally defined"—that is, when there is a claim about a person based on observations, not merely a statement about those particular observations in and of themselves. Earlier work on validity distinguished a number of varieties of validity, such as content validity, predictive validity, convergent and divergent validity, and we will say a bit more about these later. But the current view, as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, 1999) assert, is that validity is a unitary concept. Ostensibly different kinds of validity are better viewed as merely different lines of argument and different kinds of evidence for a single kind of validity. If you insist on a label for it, it would have to be *construct* validity.

Embretson (1983) distinguishes between validity arguments that concern why data gathered in a certain way ought to provide evidence about the targeted skill knowledge, and those that investigate relationships of resulting scores with other variables to support the case. These are, respectively, arguments about "construct representation" and arguments from "nomothetic span." Writing in 1983, Embretson noted that validation studies relied mainly on nomothetic arguments, using scores from assessments in their

final form or close to it. The construction of those tests, however, was guided mainly by specifications for item format and content, rather than by theoretical arguments or empirical studies regarding construct representation. The "cognitive revolution" in the latter third of the Twentieth Century provided both scientific respectability and practical tools for designing construct meaning into tests from the beginning (Embretson, 1983). The value of both lines of argument is appreciated today, with validation procedures based on nomothetic span tending to be more mature and those based on construct representation still evolving.

Reliability

Reliability concerns the adequacy of the data to support a claim, presuming the appropriateness of the warrant and the satisfactory elimination of alternative hypotheses. Even if the reasoning is sound, there may not be enough information in the data to support the claim. Later we will see how reliability is expressed quantitatively when probability-based measurement models are employed. We can mention now, though, that the procedures by which data are gathered can involve multiple steps or features that each affect the evidentiary value of data. Depending on Jim's rating of Sue's essay rather than evaluating it ourselves adds a step of reasoning to the chain, introducing the need to establish an additional warrant, examine alternative explanations, and assess the value of the resulting data.

How can we gauge the adequacy of evidence? Brennan (2000/in press) writes that the idea of repeating the measurement process has played a central role in characterizing an assessment's reliability since the work of Spearman (1904)—much as it does in physical sciences. If you weigh a stone ten times and get a slightly different answer each time, the variation among the measurements is a good index of the uncertainty associated with that measurement procedure. It is less straightforward to know just what repeating the measurement procedure means, though, if the procedure has several steps that could each be done differently (different occasions, different task, different raters), or if some of the steps can't be repeated at all (if a person learns something by working through a task, a second attempt isn't measuring the same level of knowledge). We will see that the

history of reliability is one of figuring out how to characterize the value of evidence in increasingly wider ranges of assessment situations.

Comparability

Comparability concerns the common occurrence that the specifics of data collection differ for different students, or for the same students at different times. Differing conditions raise alternative hypotheses when we need to compare students with one another or against common standards, or when we want to track students' progress over time. Are there systematic differences in the conclusions we would draw when we observe responses to Test Form A as opposed to Test Form B, for example? Or from a computerized adaptive test instead of the paper-and-pencil version? Or if we use a rating based on two judges, as opposed to the average of two, or the consensus of three? We must then extend the warrant to deal with these variations, and we must include them as alternative explanations of differences in students' scores.

Comparability overlaps with reliability, as both raise questions of how evidence obtained through one application of a data-gathering procedure might differ from evidence obtained through another application. The issue is *reliability* when we consider the two measures interchangeable—which is used is a matter of indifference to the examinee and assessor alike. Although we expect the results to differ somewhat, we don't know if one is more accurate than the other, whether one is biased toward higher values, or if they will illuminate different aspects of knowledge. The same evidentiary argument holds for both measures, and the obtained differences are what constitutes classical measurement error. The issue is *comparability* when we expect systematic differences of any of these types, but wish to compare results obtained from the two distinct processes nevertheless. A more complex evidentiary argument is required. It must address the way that observations from the two processes bear different relationships to the construct we want to measure, and it must indicate how to take these differences into account in our inferences.

Fairness

Fairness is a term that encompasses more territory than we can address in this presentation. Many of its senses concern social, political, and educational perspectives on

the uses to which assessment results inform (Willingham & Cole, 1997)—legitimate questions all, which would exist even if the chain of reasoning from observations to constructs contained no uncertainty whatsoever. Like Wiley (1991), we focus our attention here on construct meaning rather than use or consequences, and consider aspects of fairness that bear directly on this portion of the evidentiary argument.

Fairness in this sense concerns alternative explanations of assessment performances in light of other characteristics of students that we could and should take into account. Ideally, the same warrant backs inferences about many students, reasoning from their particular data to a claim about what each individually knows or can do. This is never quite truly the case in practice, for factors such as language background, instructional background, and familiarity with representations surely influence performance. When the same argument is to be applied with many students, considerations of fairness require us to examine the impact of such factors on performance and identify the ranges of their values beyond which the common warrant can no longer be justified. Drawing the usual inference from the usual data for a student who lies outside this range leads to inferential errors. If they are errors we should have foreseen and avoided, they are unfair. Ways of avoiding such errors are using additional knowledge about students to condition our interpretation of what we observe under the same procedures, and gathering data from different students in different ways, such as providing accommodations or allowing students to choose among ways of providing data (and accepting the responsibility as assessors to establish the comparability of data so obtained!)

A Framework for Assessment Design

This section lays out a schema for the evidentiary argument that underlies educational assessments, incorporating both its substantive and statistical aspects. It is based on the 'evidence–centered' framework for assessment design illustrated in Mislevy, Steinberg, and Almond (in press) and Mislevy, Steinberg, Breyer, Almond, and Johnson (1999; in press). We'll use it presently to examine psychometric principles from a more technical perspective. The framework formalizes another quotation from Messick:

> A construct-centered approach [to assessment design] 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.
>
> Messick, 1994, p. 16.

Figure 1, presented back in the introductory example, depicts elements and relationships that must be present, at least implicitly, and coordinated, at least functionally, in any assessment that has evolved to effectively serve some inferential function. Making this structure explicit helps an evaluator understand how to first gather, then reason from, data that bear on what students know and can do.

In brief, the *student model* specifies the variables in terms of which we wish to characterize students. *Task models* are schemas for ways to get data that provide evidence about them. Two components, which are links in the chain of reasoning from students' work to their knowledge and skill: The *scoring component* of the evidence model contains procedures for extracting the salient features of student's performances in task situations—i.e., observable variables—and the *measurement component* contains machinery for updating beliefs about student-model variables in light of this information. These models are discussed in more detail below. Taken together, they make explicit the evidentiary grounding of an assessment, and they guide the choice and construction of particular tasks, rubrics, statistical models, and so on. An operational assessment will generally have one student model, which may contain many variables, but may use several task and evidence models to provide data of different forms or with different rationales.

The Student Model: What complex of knowledge, skills, or other attributes should be assessed?

The values of student-model variables represent selected aspects of the infinite configurations of skill and knowledge real students have, based on a theory or a set of beliefs about skill and knowledge in the domain. These variables are the vehicle through which we determine student progress, make decisions, or plan instruction for students. The number and nature of the student model variables depend on the purpose of an assessment. A single variable characterizing overall proficiency in algebra might suffice in an assessment meant only to support a pass/fail decision; a coached practice system to help students develop the same proficiency might require a finer grained student model, to monitor how a student is doing on particular aspects of skill and knowledge for which we can offer feedback. When the purpose is program evaluation, the student model variables should reflect hypothesized ways in which a program may enjoy more or less success, or promote students' learning in some ways as opposed to others.

In the standard argument, then, *a claim about what a student knows, can do, or has accomplished is expressed in terms of values of student-model variables.* Substantive concerns about the desired outcomes of instruction, say, or the focus of a program evaluation, will suggest what the student-model variables might be, and give substantive meaning to the values of student-model variables. The student model provides a language for expressing claims about students, restricted and simplified to be sure, but one that is amenable to probability-based reasoning for drawing inferences and characterizing beliefs. A following section will explain how we can express what we know about a given student's values for these variables in terms of a probability distribution, which can be updated as new evidence arrives.

Task Models: What tasks or situations should elicit those behaviors?

A task model provides a framework for constructing and describing the situations in which examinees act. We use the term "task" in the sense proposed by Haertel and Wiley (1993), to refer to a "goal-directed human activity to be pursued in a specified manner, context, or circumstance." A task can thus be an open-ended problem in a computerized

simulation, a long-term project such as a term paper, a language-proficiency interview, or a familiar multiple-choice or short-answer question.

A task model specifies the environment in which the student will say, do, or produce something; for example, characteristics of stimulus material, instructions, help, tools, and so on. It also specifies the work product, or the form in which what the student says, does, or produces will be captured. But again it is substantive theory and experience that determine the kinds of situations can evoke behaviors that provide clues about the targeted knowledge and skill, and the forms in which those clues can be expressed and captured.

To create a particular task, an assessment designer has explicitly or implicitly assigned specific values to task model variables, provided materials that suit the specifications there given, and set the conditions that are required to interact with the student. A task thus describes particular circumstances meant to provide the examinee an opportunity to act in ways that produce evidence about what they know or can do more generally. For a particular task, *the values of its task model variables constitute data for the evidentiary argument, characterizing the situation in which the student is saying, doing, or making something.*

It is useful to distinguish task models from the scoring models discussed in the next section, as the latter concern what to attend to in the resulting performance and how to evaluate what we see. Distinct and possibly quite different evaluation rules could be applied to the same work product from a given task. Distinct and possibly quite different student models, designed to serve different purposes or derived from different conceptions of proficiency, could be informed by performances on the same tasks. The substantive arguments for the evidentiary value of behavior in the task situation will overlap in these cases, but the specifics of the claims and thus the specifics of the statistical links in the chain of reasoning will differ.

Evidence Models: What behaviors or performances should reveal the student constructs, and what is the connection?

An evidence model lays out the part of the evidentiary argument that concerns reasoning from the observations in a given task situation to revising beliefs about student model variables. Figure 1 shows there are two parts to the evidence model.

The *scoring component* contains "evidence rules" for extracting the salient features of whatever the student says, does, or creates in the task situation—i.e., the "work product" that is represented by the jumble of shapes in the rectangle at the far right of the evidence model. A work product is a unique human production, perhaps as simple as a response to a multiple-choice item, or as complex as repeated cycles of treating and evaluating patients in a medical simulation. The squares coming out of the work product represent "observable variables," or evaluative summaries of what the assessment designer has determined are the key aspects of the performance (as captured in one or more work products) to serve the assessment's purpose. Different aspects could be captured for different purposes. For example, a short impromptu speech contains information about a student's subject matter knowledge, presentation capabilities, or English language proficiency; any of these, or any combination, could be the basis of one or more observable variables. As a facet of fairness, however, the student should be informed of which aspects of her performance are being evaluated, and by what criteria. For students failing to understand how their work will be scored is an alternative hypothesis for poor performance we can and should avoid.

Scoring rules map unique performances into a common interpretative framework, thus laying out what is important in a performance. These rules can be as simple as determining whether the response to a multiple-choice item is correct, or as complex as an expert's holistic evaluation of multiple aspects of an unconstrained patient-management solution. They can be automated, demand human judgment, or require both in combination. *Values of the observable variables describe properties of the particular things a student says, does, or makes. As such, they constitute data about what the student knows, can do, or has accomplished as more generally construed* in the standard argument.

It is important to note that substantive concerns drive the definition of observable variables. Statistical analyses can be used to refine definitions, compare alternatives, or improve data-gathering procedures, again looking for patterns that call a scoring rule into question. But it is the conception of what to observe that concerns validity directly, and raises questions of alternative explanations that bear on comparability and fairness.

The *measurement component* of the Evidence Model tells how the observable variables depend, in probabilistic terms, on student model variables, another essential link in the evidentiary argument. This is the foundation for the reasoning that is needed to synthesize evidence across multiple tasks or from different performances. Figure 1 shows how the observables are modeled as depending on some subset of the student model variables. The familiar models from test theory that we discuss in a following section, including classical test theory and item response theory, are examples. We can adapt these ideas to suit the nature of the student model and observable variables in any given application (Almond & Mislevy, 1999). Again, substantive considerations must underlie why these posited relationships should hold; the measurement model formalizes the patterns they imply.

It is a defining characteristic of psychometrics to model observable variables as probabilistic functions of unobservable student variables. The measurement model is almost always a probability model. The probability-based framework model may extend to the scoring model as well, as when judgments are required to ascertain the values of observable variables from complex performances. Questions of accuracy, agreement, leniency, and optimal design arise, and can be addressed with a measurement model that addresses the rating link as well as the synthesis link in the chain of reasoning. The generalizability and rater models discussed below are examples of this.

Psychometric Principles and Probability-Based Reasoning

The role of probability-based reasoning in the assessment

This section looks more closely at what is perhaps the most distinctive characteristic of psychometrics, namely, the use of statistical models. Measurement models are a particular form of reasoning from evidence; they provide explicit, formal rules for how to integrate the many pieces of information that may be relevant to a particular inference

about what students know and can do. Statistical modeling, probability-based reasoning more generally, is an approach to solving the problem of "reverse reasoning" through a warrant, from particular data to a particular claim. Just how can we reason from data to claim for a particular student, using a measurement model established for general circumstances--usually far less than certain, typically with qualifications, perhaps requiring side conditions we don't know are satisfied? How can we synthesize the evidentiary value of multiple observations, perhaps from different sources, often in conflict?

The essential idea is to approximate the important substantive relationships in some real world problem in terms of relationships among variables in a probability model. A simplified picture of the real-world situation results. A useful model does not explain all the details of actual data, but it does capture the significant patterns among them. What is important is that in the space of the model, the machinery of probability-based reasoning indicates exactly how reverse reasoning is to be carried out (specifically, through Bayes theorem), and how different kinds and amounts of data should affect our beliefs. The trick is to build a probability model that both captures the important real-world patterns and suits the purposes of the problem at hand.

Measurement models concern the relationships between students' knowledge and their behavior. A student is modeled in terms of variables (θ) that represent the facets of skill or knowledge that suit the purpose of the assessment, and the data (X) are values of variables that characterize aspects of the observable behavior. We posit that the student model variables account for observable variables in the following sense: We don't know exactly what any student will do on a particular task, but for people with any given value of θ, there is a probability distribution of possible values of X, say $p(X|\theta)$. This is a mathematical expression of what we might expect to see in data, given any possible values of student-model variables. The way is open for reverse reasoning, from observed Xs to likely θs, as long as different values of θ produce different probability distributions for X. We don't know the values of the student-model variables in practice; we observe "noisy" data presumed to have been determined by them, and through the probability model reason back to what their values are likely to be.

Choosing to manage information and uncertainty with probability-based reasoning, with its numerical expressions of belief in terms of probability distributions, does not constrain one to any particular forms of evidence or psychological frameworks. That is, it says nothing about the number or nature of elements of X, or about the character of the performances, or about the conditions under which performances are produced. And it says nothing about the number or nature of elements of θ, such as whether they are number values in a differential psychology model, production-rule mastery in a cognitive model, or tendencies to use resources effectively in a situative model. In particular, using probability-based reasoning does not commit us to long tests, discrete tasks, or large samples of students. For example, probability-based models have been found useful in modeling patterns of judges' ratings in the previously mentioned Advanced Placement Portfolio Art assessment (Myford & Mislevy, 1995), about as open-ended as large-scale, high-stakes educational assessments get, and in modeling individual students' use of production rules in a tutoring system for solving physics problems (Martin & vanLehn, 1995).

Now a measurement model in any case is not intended to account for every detail of data; it is only meant to approximate the important patterns. The statistical concept of *conditional independence* formalizes the working assumption that if the values of the student model variables were known, there would have no further information in the details. The fact that every detail of a student's responses could in principle contain information about what a student knows or how she is thinking underscores the constructive and purposive nature of modeling. We use a model at a given grainsize or with certain kinds of variables not because we think that is somehow "true", but rather because it adequately expresses the patterns in the data in light of the purpose of the assessment. Adequacy in a given application depends on validity, reliability, comparability, and fairness in ways we shall discuss further, but characterized in ways and demanded in degrees that depend on that application: The purpose of the assessment, the resources that are available, the constraints that must be accommodated. We might model the same troubleshooting performances in terms of individual problem steps for an intelligent tutoring system, in terms of general areas of strength and weakness for a

diagnostic assessment, and simply in terms of overall success rate for a pass/fail certification test.

Never fully believing the statistical model we are reasoning through, we bear the responsibility of assessing model fit, in terms of both persons and items. We must examine the ways and the extent to which the real data depart from the patterns in the data, calling our attention to failures of conditional independence—places where our simplifying assumptions miss relationships that are surely systematic, and possibly important, in the data. Finding substantial misfit causes us to re-examine the arguments that tell us what to observe and how to evaluate it.

Probability-based reasoning in classical test theory

This section illustrates the ideas from the preceding discussion in the context of Classical Test Theory (CTT). In CTT, the student model is represented as a single continuous unobservable variable, the true score θ. The measurement model simply tells us to think of an observed score X as the true score plus an error term. If a CTT measurement model were used in the BEAR example, it would address the sum of the student scores on a set of assessment tasks as the observed score.

Figure 9 pictures the situation, in a case that concerns Sue's (unobservable) true score and her three observed scores on parallel forms of the same test; that is, they are equivalent measures of the same construct, and have the same means and variances. The probability distribution p(θ) expresses our belief about Sue's θ before we observe her test scores, the Xs. The conditional distributions p($X_j|\theta$)[ii] indicate the probabilities of observing different values of X_j if θ took any given particular value. Modeling the distribution of each X_j to depend on θ but not the other Xs is an instance of conditional independence; more formally, we write p(X_1, X_2, $X_3|\theta$)= p($X_1|\theta$) p($X_2|\theta$) p($X_3|\theta$). Under CTT we may obtain a form for the p($X_j|\theta$)s by proposing that

$$X_j = \theta + E_j, \qquad (1)$$

where E_j is an "error" term, normally-distributed with a mean of zero and a variance of σ_E^2.[iii] Thus $X_j|\theta \sim N(\theta, \sigma_E)$. This statistical structure quantifies the patterns that the substantive arguments express qualitatively, in a way that tells us exactly how to carry out reverse reasoning for particular cases. If $p(\theta)$ expresses belief about Sue's θ *prior* to observing her responses, belief *posterior* to learning them is denoted as $p(\theta|x_1,x_2,x_3)$ and is calculated by Bayes theorem as

$$p(\theta|x_1,x_2,x_3) \propto p(\theta)\, p(x_1|\theta)\, p(x_2|\theta)\, p(x_3|\theta).$$

(The lower case x's here denote particular values of X's.)

[[Figure 9—CTT DAG—about here]]

Figure 10 gives the numerical details for a hypothetical example, calculated with a variation of an important early result called Kelley's formula for estimating true scores (Kelley, 1927). Suppose that from a large number of students like Sue, we've estimated that the measurement error variance is $\sigma_E^2 = 25$, and for the population of students, θ follows a normal distribution with mean 50 and standard deviation 10. We now observe Sue's three scores, which take the values 70, 75, and 85. We see that the posterior distribution for Sue's θ is a normal distribution with mean 74.6 and standard deviation 2.8.

[[Figure 10—CTT numerical example—about here]]

The additional backing that was used to bring the probability model into the evidentiary argument was an analysis of data from students like Sue. Spearman's (1904) seminal insight was that if their structure is set up in the right way[iv], it is possible to estimate the quantitative features of relationships like this, among both variables that could be observed and others which by their nature never can be. The index of measurement accuracy in CTT is the *reliability coefficient* ρ, which is the proportion of variance in observed scores in a population of interest that is attributable to true scores as

opposed to the total variance, which is composed of true score variance and noise. It is defined as follows:

$$\rho = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_E^2}, \tag{2}$$

where σ_θ^2 is the variance of true score in the population of examinees and σ_E^2 is the variance of the error components—neither of which is directly observable! With a bit of algebra, though, Spearman demonstrated that if Equation 1 holds, correlations among pairs of *X*s will approximate ρ. We may then estimate the contributions of true score and error, or σ_θ^2 and σ_E^2, as proportions ρ and (1-ρ) respectively of the observed score variance. The intuitively plausible notion is that correlations among exchangeable measures of the same construct tell us how much to trust comparisons among examinees from a single measurement. (As an index of measurement accuracy, however, ρ suffers from its dependence on the variation among examinees' true scores as well as on the measurement error variance of the test. For a group of examinees with no true-score variance, the reliability coefficient is zero no matter how much evidence a test provides about each of them. We'll see how item response theory extends the idea of measurement accuracy.)

What's more, patterns among the observables can be so contrary to those the model would predict that we suspect the model isn't right. Sue's values of 70, 75, and 85 are not identical, but neither are they surprising as a set of scores (Figure 10 calculates a chi-squared index of fit for Sue). Some students have higher scores than Sue, some have lower scores, but the amount of variation within a typical student's set of scores is in this neighborhood. But Richard's three scores of 70, 75, and 10 *are* surprising. His high fit statistic (a chi-square of 105 with 2 degrees of freedom) says his pattern is very unlikely from parallel tests with an error variance of 25 (less than one in a billion). Richard's responses are so discordant with a statistical model that expresses patterns under the standard argument that we suspect that the standard argument does not apply. We must go beyond the standard argument to understand what has happened, to facts the standard data

do not convey. Our first clue is that his third score is particularly different from both the other two and from the prior distribution.

Classical test theory's simple model for examinee characteristics suffices when one is just interested in is a single aspect of student achievement, tests are only considered as a whole, and all students take tests that are identical or practically so. But the assumptions of CTT have generated a vast armamentarium of concepts and tools that help the practitioner examine the extent to which psychometric principles are being attained in situations when the assumptions are adequate. These tools include reliability indices that can be calculated from multiple items in a single test, formulas for errors of measuring individual students, strategies for selecting optimal composites of tests, formulas for approximating how long a test should be to reach a required accuracy, and methods for equating tests. The practitioner working in situations that CTT encompasses will find a wealth of useful formulas and techniques in Gulliksen's (1950/1987) classic text.

The advantages of using probability-based reasoning in assessment

Because of its history, the very term "psychometrics" connotes a fusion of the inferential logic underlying Spearman's reasoning with his psychology (trait psychology, in particular with intelligence as an inherited and stable characteristic) and his data-gathering methods (many short, 'objectively-scored,' largely decontextualized tasks). The connection, while historically grounded, is logically spurious, however. For the kinds of problems that CTT grew to solve are not just Spearman's problems, but ones that ought to concern anybody who is responsible for making decisions about students, evaluating the effects of instruction, or spending scarce educational resources—whether or not Spearman's psychology or methodology are relevant to the problem at hand.

And indeed, the course of development of test theory over the past century has been to continually extended the range of problems to which this inferential approach can be applied—to claims cast in terms of behavioral, cognitive, or situative psychology[v]; to data that may be embedded in context, require sophisticated evaluations, or address multiple interrelated aspects of complex activities. We will look at some of these developments in the next section. But it is at a higher level of abstraction that psychometric principles are

best understood, even though it is with particular models and indices that they are investigated in practice.

When it comes to examining psychometric properties, embedding the assessment argument in a probability model offers the following advantages:

1. Using the calculus of probability-based reasoning, once we ascertain the values of the variables in the data, we can express our beliefs about the likely values of the student estimates in terms of probability distributions—given that the model is both generally credible (#3 below) and applicable to the case at hand (#4 below).
2. The machinery of probability-based reasoning is rich enough to handle many recurring challenges in assessment, such as synthesizing information across multiple tasks, characterizing the evidentiary importance of elements or assemblages of data, assessing comparability across different bodies of evidence, and exploring the implications of judgment, including different numbers and configurations of raters.
3. Global model-criticism techniques allow us to not only fit models to data, but to determine where and how the data do not accord well with the models. Substantive considerations suggest the structure of the evidentiary argument; statistical analyses of ensuing data through the lens of a mathematical model help us assess whether the argument matches up with what we actually see in the world. For instance, detecting an unexpected interaction between performance on an item and students' cultural backgrounds alerts us to an alternate explanation of poor performance. We are then moved to improve the data gathering methods, constrain the range of use, or rethink the substantive argument.
4. Local model-criticism techniques allow us to monitor the operation of the reverse-reasoning step for individual students even after the argument, data-collection methods, and statistical model are up and running. Patterns of observations that are unlikely under the common argument can be flagged (e.g., Richard's high chi-square value), thus avoiding certain unsupportable inferences and drawing attention to those cases that call for additional exploration.

Implications for psychometric principles

Validity

Some of the historical "flavors" of validity are statistical in nature. Predictive validity is the degree to which scores in selection tests correlate with future performance. Convergent validity looks for high correlations of a test's scores with other sources of evidence about the targeted knowledge and skills, while divergent validity looks for low correlations with evidence about irrelevant factors (Campbell & Fiske, 1959). Concurrent validity examines correlations with other tests presumed to provide evidence about the same or similar knowledge and skills.

The idea is that substantive considerations that justify an assessment's conception and construction can be put to empirical tests. In each of the cases mentioned above, relationships are posited among observable phenomena that would hold if the substantive argument were correct, and see if in fact they do; that is, exploring the nomothetic net. These are all potential sources of backing for arguments for interpreting and using test results, and they are at the same time explorations of plausible alternative explanations.

Consider, for example, assessments meant to support decisions about whether a student has attained some criterion of performance (Ercikan & Julian, 2001, Hambleton & Slater, 1997). These decisions, typically reported as proficiency or performance level scores, which are increasingly being considered to be useful in communicating assessment results to students, parents and the public as well as for evaluation of programs, involve classification of examinee performance to a set of proficiency levels. Rarely do the tasks on such a test exhaust the full range of performances and situations users are interested in. Examining the validity of a proficiency test from this nomothetic-net perspective would involve seeing whether students who do well on that test also perform well in more extensive assessment, obtain high ratings from teachers or employers, or succeed in subsequent training or job performance.

Statistical analyses of these kinds have always been important after the fact, as significance-focused validity studies informed, constrained, and evaluated the use of a test—but they rarely prompted more than minor modifications to its contents. Rather, Embretson (1998) notes, substantive considerations have traditionally driven assessment

construction. Neither of two meaning-focused lines of justification that were considered forms of validity used probability-based reasoning. They were content validity, which concerned the nature and mix of items in a test, and face validity, which is what a test appears to be measuring on the surface, especially to non-technical audiences. We will see in our discussion of item response theory how statistical machinery is increasingly being used in the exploration of construct representation as well in after-the-fact validity studies.

Reliability

Reliability, historically, was used to quantify the amount of variation in test scores that reflected 'true' differences among students, as opposed to noise (Equation 2). The correlations between parallel tests forms we used in classical test theory are one way to estimate reliability in this sense. Internal consistency among test items, as gauged by the KR-20 formula (Kuder & Richardson, 1937) or Cronbach's (1951) Alpha coefficient, is another. A contemporary view sees reliability as the evidentiary value that a given realized or prospective body of data would provide for a claim—more specifically, the amount of information for revising belief about an inference involving student-model variables, be it an estimate for a given student, a comparison among students, or a determination of whether a student has attained some criterion of performance.

A wide variety of specific indices or parameters can be used to characterize evidentiary value. Carrying out a measurement procedure two or more times with supposedly equivalent alternative tasks and raters will not only ground an estimate of its accuracy, as in Spearman's original procedures, but it demonstrates convincingly that there is some uncertainty to deal with in the first place (Brennan, 2000/in press). The KR-20 and Cronbach's alpha apply the idea of replication to tests that consist of multiple items, by treating subsets of the items as repeated measures. These CTT indices of reliability appropriately characterize the amount of evidence for comparing students in a particular population with one another, but not necessarily for comparing them against a fixed standard, or for comparisons in other populations, or for purposes of evaluating schools or instructional programs. In this sense CTT indices of reliability are tied to particular populations and inferences.

Since reasoning about reliability takes place in the realm of the measurement model (assuming that it is both correct and appropriate), it is possible to approximate the evidentiary value of not only the data in hand, but the value of similar data gathered in somewhat different ways. Under CTT, the Spearman-Brown formula (Brown, 1910; Spearman, 1910) can be used to approximate the reliability coefficient that would result from doubling the length of a test:

$$\rho_{\text{double}} = \frac{2\rho}{1+\rho}. \tag{3}$$

That is, if ρ is the reliability of the original test, then ρ_{double} is the reliability of an otherwise comparable test with twice as many items. Empirical checks have shown that these predictions can hold up quite well—but not if the additional items differ as to their content or difficulty, or if the new test is long enough to fatigue students. In these cases, the real-world counterparts of the modeled relationships are stretched so far that the results of reasoning through the model fail.

Extending this thinking to a wider range of inferences, generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) permits predictions for the accuracy of similar tests with different numbers and configurations of raters, items, and so on. And once the parameters of tasks have been estimated under an item response theory (IRT) model, one can even assemble tests item by item to individual examinees on the fly, to maximize the accuracy with which each is assessed. (Later we'll point to some "how-to" references for g-theory and IRT.)

Typical measures of accuracy used in CTT are not sufficient for examining accuracy of the decisions concerning criterion of performance discussed above. In CTT framework, the classification accuracy is defined as the extent to which classification of students based on their observed test scores agree with those based on their true scores (Traub & Rowley, 1980). One of the two commonly used measures of classification accuracy is a simple measure of agreement, $p_{0,}$ defined as

$$p_0 = \sum_{l=1}^{L} p_{ll},$$

where p_{ll} represents the proportion of examinees who were classified into the same proficiency level (l=2,..,5) according to their true score and observed score. The second is Cohen's κ coefficient (Cohen, 1960). This statistic is similar to the proportion agreement p_0 , except that it is corrected for the agreement which is due to chance. The coefficient is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c},$$

where

$$p_c = \sum_{l=1}^{L} p_{l.} p_{.l}.$$

The accuracy of classifications based on test scores are critically dependent on measurement accuracy at the cut-score points (Ercikan & Julian, 2001; Hambleton & Slater, 1997). Even though higher measurement accuracy tends to imply higher classification accuracy, higher reliability such as one indicated by KR-20 or Coefficient alpha does not imply higher classification accuracy. These measures provide an overall indication of measurement accuracy provided by the test for all examinees, however, they do not provide information about the measurement accuracy provided at the cut-scores. Therefore, they are not sufficient indicators of accuracy of classification decisions made based on test performance.

On the other hand, measurement accuracy is expected to vary for different score ranges resulting in variation in classification accuracy. This points to a serious limitation of interpretability of single indices that are intended to represent classification accuracy of a test given a set of cut-scores. Ercikan & Julian (2001) study found that the classification accuracy can be dramatically different for examinees at different ability levels. Their results demonstrated that comparing classification accuracy across tests could be deceptive, since classification accuracy may be higher for one test for certain score ranges and lower for others. Based on these limitations of interpretability of classification accuracy for different score ranges, these authors recommend that classification accuracy be reported separately for different score ranges.

Comparability

Comparability, it will be recalled, concerns the equivalence of inference when different bodies of data are gathered to compare students, or to assess change from the same students at different points in time. Within a statistical framework, we can build models that address quantitative aspects of questions such as these: Do the different bodies of data have such different properties as evidence as to jeopardize the inferences? Are conclusions about students' knowledge biased in one direction or another when different data are gathered? Is there more or less weight for various claims under the different alternatives?

A time-honored way of establishing comparability has been creating parallel test forms. A common rationale is developed to create collections of tasks which, taken together, can be argued to provide data about the same targeted skills and knowledge—differing, it is hoped, only in incidentals that do not accumulate. Defining a knowledge-by-skills matrix, for example, writing items in each cell, and constructing tests by selecting the same numbers of tasks from each cell for every test form. The same substantive backing thus grounds all the forms.

But it would be premature to presume that equal scores from these tests constitute equivalent evidence about students' knowledge. Despite care in their construction, possible differences between the tests as to difficulty or amount of information must be considered as an alternative explanation for differing performances among students. Empirical studies and statistical analyses enter the picture at this point, in the form of equating studies (Petersen, Kolen, & Hoover, 1989). Finding that similar groups of students systematically perform better on Form A than on Form B confirms the alternative explanation. Adjusting scores for Form B upward to match the resulting distributions addresses this concern, refines the chain of reasoning to take form differences into account when drawing claims about students, and enters the compendium of backing for the assessment system as a whole. Below we shall see that IRT extends comparability arguments to test forms that differ in difficulty and accuracy, if they can satisfy the requirements of a more ambitious statistical model.

Fairness

The meaning-focused sense of fairness we have chosen to highlight concerns a claim that would follow from the common argument, but would be called into question by an alternative explanation sparked by other information we could and should have taken into account. When we extend the discussion of fairness to statistical models, we find macro-level and micro-level strategies to address this concern.

Macro-level strategies of fairness fall within the broad category of what the assessment literature calls validity studies, and are investigations in the nomothetic net. They address broad patterns in test data, at the level of arguments or alternative explanations in the common arguments that are used with many students. Suppose the plan for the assessment is to use data (say, essay responses) to back a claim about a student's knowledge (e. g., the student can back up opinions with examples) through an argument (e. g., students in a pretest who are known to be able to do this in their first language are observed to do so in essays that ask for them to), without regard to a background factor (such as a students' first language). The idea is to gather, from a group of students, data that include their test performances but also include information about their first language and higher-quality validation data about the claim (e.g., interviews in the students' native languages). The empirical question is whether inferences from the usual data to the claim (independently evidenced by the validity data) differ systematically with first language. In particular, are there students who can back arguments with specifics in their native language, but fail to do so on the essay test because of language difficulties? If so, the door is open to distorted inferences about argumentation skills for limited-English speakers, if one proceeds from the usual data through the usual argument, disregarding language proficiency.

What can we do when the answer is "yes"? Possibilities include improving the data collected for all students, taking their first language into account when reasoning from data to claim (recognizing the language difficulties can account for poor performance even when the skill of interest is present), and pre-identifying students whose limited language proficiencies are likely to lead to flawed inferences about the targeted

knowledge. In this last instance, additional or different data could be used for these students, such as an interview or an essay in their primary language.

These issues are particularly important in assessments used for making consequential proficiency-based decisions, in ways related to the points we raised concerning the validity of such tests. Unfair decisions are rendered if (a) alternative valid means of gathering data for evaluating proficiency yield results that differ systematically from the standard assessment, and (b) the reason can be traced to requirements for knowledge or skills (e.g., proficiency with the English language) that are not central to the knowledge or skill that is at issue (e.g., constructing and backing an argument).

The same kinds of investigations can be carried out with individual tasks as well as with assessments as a whole. One variation on this theme can be used with assessments that are composed of several tasks, to determine whether individual tasks interact with first language in atypical ways. These are called studies of DIF, or differential item functioning (Holland & Wainer, 1993).

Statistical tools can also be used to implement micro-level strategies to call attention to cases in which a routine application of the standard argument could produce a distorted and possibly unfair inference. The common argument provides a warrant to reason from data to claim, with attendant caveats for unfairness associated with factors (such as first language) that have been dealt with at the macro level. But the argument may not hold for some individual students for other reasons, which have not yet been dealt with at the macro level, perhaps could not have been anticipated at all. Measurement models characterize patterns in students' data that are typical if the general argument holds. Patterns that are unlikely can signal that the argument may not apply with a given student on a given assessment occasion. Under IRT, for example, 'student misfit' indices take high values for students who miss items that are generally easy while correctly answering ones that are generally hard (Levine & Drasgow, 1982.)

Some Other Widely-Used Measurement Models

The tools of classical test theory have been continually extended and refined in the time since Spearman, to the extensive toolkit by Gulliksen (1950/1987), and to the sophisticated theoretical framework laid out in Lord and Novick (1968). Lord and

Novick aptly titled their volume *Statistical theories of mental test scores*, underscoring their focus on the probabilistic reasoning aspects in the measurement-model links of the assessment argument—not the purposes, not the substantive aspects, not the evaluation rules that produce the data. Models that extend the same fundamental reasoning for this portion of assessment arguments to wider varieties of data and student models include generalizability theory, item response theory, latent class models, and multivariate models.

Each of these extensions offers more options for characterizing students and collecting data in a way that can be embedding in a probability model. The models do not concern themselves directly with substantive aspects of an assessment argument, but substantive considerations often have much to say about how one should think about students' knowledge, and what observations should contain evidence about it. The more measurement models that are available and the more kinds of data than can be handled, the better assessors can match rigorous models with the patterns their theories and their needs concern. This bolsters evidentiary arguments (validity), extends quantitative indices of accuracy to more situations (reliability), enables more flexibility in observational settings (comparability), and enhances the prospects of detecting students whose data are at odds with standard argument (fairness).

Generalizability Theory

Generalizability Theory (g-theory) extends classical test theory by allowing us to examine how different aspects of the observational setting affect the evidentiary value of test scores. As in CTT, the student is characterized by overall proficiency in some domain of tasks. However, the measurement model can now include parameters that correspond to "facets" of the observational situation such as features of tasks (i.e., task-model variables), numbers and designs of raters, and qualities of performance that will be evaluated. An observed score of a student in a generalizability study of an assessment consisting of different item types and judgmental scores is an elaboration of the basic CTT equation:

$$X_{ijk} = \theta_i + \tau_j + \varsigma_k + E_{ijk},$$

where we now address the observed score is from Examinee *i*, to Item-Type *j*, as evaluated by Rater *k*; θ_i is the true score of Examinee *i*; and τ_j and ς_k are, respectively, effects attributable to Item-Type *j* and Rater *k*.

Researchers carry out a generalizability study, or g-study, to estimate the amount of variation associated with different facets. The accuracy of estimation of scores for a given configuration of tasks can be calculated from these variance components, the numbers of items and raters, and the design in which data are collected. A "generalizability coefficient" is an extension of the CTT reliability coefficient: it is the proportion of true variance among students for the condition one wishes to measure, divided by the variance among observed scores among the measurements that would be obtained among repeated applications of the measurement procedure that is specified (how many observations, fixed or randomly selected; how many raters rating each observation, different or same raters for different items, etc.). If, in the example above, we wanted to estimate θ using one randomly selected item, scored as the average of the ratings from two randomly selected raters, the coefficient of generalizability, denoted here as α, would be calculated as follows:

$$\alpha = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\tau^2 + \left(\sigma_\varsigma^2 + \sigma_E^2\right)/2},$$

where σ_θ^2, σ_τ^2, σ_ς^2, and σ_E^2 are variance coefficients for examinees, item-types, raters, and error respectively.

The information resulting from a generalizability study can thus guide decisions about how to design procedures for making observations; for example, what design for assigning raters to performances, how many tasks and raters, and whether to average across raters, tasks, etc. In the BEAR Assessment System, a g-study could be carried out to see which type of assessment, embedded tasks or link items, resulted in more reliable scores. It could also be used to examine whether teachers were as consistent as external raters.

G-theory offers two important practical advantages over CTT: First, generalizability models allow us to characterize how the particulars of the evaluation rules and task model variables affect the value of the evidence we gain about the student for various inferences. Second, this information is expressed in terms that allow us to project these evidentiary-value considerations to designs we have not actually used, but which could be constructed from elements similar to the ones we have observed. G-theory thus provides far-reaching extensions of the Spearman-Brown formula (Equation 3), for exploring issues of reliability and comparability in a broader array of data-collection designs than CTT can.

Generalizability theory was developed by Professor Lee Cronbach and his colleagues, and their monograph *The dependability of behavioral measurements* (Cronbach et al., 1972) remains a valuable source of information and insight. More recent sources such as Shavelson and Webb (1991) and Brennan (1983) provide the practitioner with friendlier notation and examples to build on.

Item Response Theory (IRT)

Classical test theory and generalizability theory share a serious shortcoming: measures of examinees are confounded with the characteristics of test items. It is hard to compare examinees who have taken tests that differ by even as much as a single item, or to compare items that have been administered to different groups of examinees. Item Response Theory (IRT) was developed to address this shortcoming. In addition, IRT can be used to make predictions about test properties using item properties and to manipulate parts of tests to achieve targeted measurement properties. Hambleton (1993) gives a readable introduction to IRT, while van der Linden and Hambleton (1997) provide a comprehensive though technical compendium of current IRT models. IRT further extended probability-based reasoning for addressing psychometric principles, and it sets the stage for further developments. We'll start with a brief overview of the key ideas.

At first, the student model under IRT seems to be the same as it is under CTT and g-theory, namely, single variable measuring students' overall proficiency in some domain of tasks. Again the statistical model does not address the nature of that proficiency. The structure of the probability-based portion of the argument is the same as shown in Figure 9: conditional independence among observations given an underlying, inherently

unobservable, proficiency variable θ. But now the observations are responses to individual tasks. For Item j, the IRT model expresses the probability of a given response x_j as a function of θ and parameters β_j that characterize Item j (such as its difficulty):

$$f(x_j;\theta,\beta_j). \tag{4}$$

Under the Rasch (1960/1980) model for dichotomous (right/wrong) items, for example, the probability of a correct response takes the following form:

$$\text{Prob}(X_{ij}=1|\theta_i,\beta_j) = f(1;\theta_i,\beta_j) = \Psi(\theta_i - \beta_j), \tag{5}$$

where X_{ij} is the response of Student i to Item j, 1 if right and 0 if wrong; θ_i is the proficiency parameter of Student i; β_j is the difficulty parameter of Item j; and $\Psi(\cdot)$ is the logistic function, $\Psi(x) = \exp(x)/[1+\exp(x)]$. The probability of an incorrect response is then

$$\text{Prob}(X_{ij}=0|\theta_i,\beta_j) = f(0;\theta_i,\beta_j) = 1-\Psi(\theta_i - \beta_j). \tag{6}$$

Taken together, Equations 5 and 6 specify a particular form for the item response function, Equation 4. Figure 11 depicts Rasch item response curves for two items, Item 1 an easy one, with β_1=-1 and Item 2 a hard one with β_2=2. It shows the probability of a correct response to each of the items for different values of θ. For both items, the probability of a correct response increases toward one as θ increases. Conditional independence means that for a given value of θ, the probability of Student i making responses x_{i1} and x_{i2} to the two items is the product of terms like Equations 5 and 6:

$$\text{Prob}(X_{i1}=x_{i1}, X_{i2}=x_{i2}|\theta_i,\beta_1,\beta_2) = \text{Prob}(X_{i1}=x_{i1}|\theta_i,\beta_1)\ \text{Prob}(X_{i2}=x_{i2}|\theta_i,\beta_2). \tag{7}$$

[[Figure 11—two item response curves]]

All this is reasoning *from* model and given parameters, *to* probabilities of not-yet-observed responses; as such, it is part of the warrant in the assessment argument, to be backed by empirical estimates and model criticism. In applications we need to reason in the reverse direction. Item parameters will have been estimated and responses observed, and we need to reason *from* an examinee's *x*s, *to* the value of θ. Equation 7 is then calculated as a function of θ with x_{i1} and x_{i2} fixed at their observed values; this is the likelihood function. Figure 12 shows the likelihood function that corresponds to X_{i1}=0 and X_{i2}=1. One can estimate θ by the point at which the likelihood attains its maximum (around .75 in this example), or use Bayes theorem to combine the likelihood function with a prior distribution for θ, $p(\theta)$, to obtain the posterior distribution $p(\theta|x_{i1},x_{i2})$.

[[Figure 12—a likelihood function]]

The amount of information about θ available from Item j, $I_j(\theta)$, can be calculated as a function of θ, β_j, and the functional form of f (see the references mentioned above for formulas for particular IRT models). Under IRT, the amount of information for measuring proficiency at each point along the scale is simply the sum of these item-by-item information functions. The square root of the reciprocal of this value is the standard error of estimation, or the standard deviation of estimates of θ around its true value. Figure 13 is the test information curve that corresponds to the two items in the preceding example. It is of particular importance in IRT that once item parameters have been estimated ("calibrating" them), estimating individual students' θs and calculating the accuracy of those estimates can be accomplished for any subset of items. Easy items can be administered to fourth graders and harder ones to fifth graders, for example, but all scores arrive on the same θ scale. Different test forms can be given as pretests and posttests, and differences of difficulty and accuracy are taken into account.

[[Figure 13—an information function]]

IRT helps assessors achieve psychometric quality in several ways.

Concerning validity: The statistical framework indicates the patterns of observable responses that would occur in data if it were actually the case that a single underlying proficiency did account for all the systematic variation among students and items. All the tools of model criticism from five centuries of probability-based reasoning can be brought to bear to assess how well an IRT model fits a given data set, and where it breaks down, now item by item, student by student. The IRT model does not address the substance of the tasks, but by highlighting tasks that are operating differently than others, or proving harder or easier than expected, it helps test designers improve their work.

Concerning reliability: Once item parameters have been estimated, a researcher can gauge the precision of measurement that would result from different configurations of tasks. Precision of estimation can be gauged uniquely for any matchup between a person and a set of items. We are no longer bound to measures of reliability that are tied to specific populations and fixed test forms.

Concerning comparability: IRT offers strategies beyond the reach of CTT and g-theory for assembling tests that "measure the same thing." These strategies capitalize on the above-mentioned capability to predetermine the precision of estimation from different sets of items at different levels of θ. Tests that provide optimal measurement for mastery decisions can be designed, for example, or tests that provide targeted amounts of precision at specified levels of proficiency (van der Linden, 1998). Large content domains can be covered in educational surveys by giving each student only a sample of the tasks, yet using IRT to map all performances onto the same scale. The National Assessment of Educational Progress, for example, has made good use of the efficiencies of this item sampling in conjunction with reporting based on IRT (Messick, Beaton, & Lord, 1983). Tests can even be assembled on the fly in light of a student's previous responses as assessment proceeds, a technique called adaptive testing (see Wainer et al., 2000, for practical advice on constructing computerized adaptive tests).

Concerning fairness: An approach called "differential item functioning" (DIF) analysis, based on IRT and related methods, has enabled both researchers and large–scale assessors to routinely and rigorously test for a particular kind of unfairness (e.g., Holland and Thayer, 1988; Lord, 1980). The idea is that a test score, such as a number-correct or

an IRT θ estimate, is a summary over a large number of item responses, and a comparison of students at the level of scores implies that they are similarly comparable across the domain being assessed. But what if some items of the items are systematically harder for students from a group defined by cultural or educational background, for reasons that are not related to the knowledge or skill that is meant to be measured? This is DIF, and it can be formally represented in a model containing interaction terms for items by groups by overall proficiency—an interaction whose presence can threaten score meaning and distort comparisons across groups. A finding of significant DIF can imply that the observation framework needs to be modified, or if the DIF is common to many items, that the construct-representation argument is oversimplified.

DIF methods have been used in examining differential response patterns for gender and ethnic groups for the last two decades and for language groups more recently. They are now being used to investigate whether different groups of examinees of approximately the same ability appear to be using differing cognitive processes to respond to test items. Such uses include examining whether differential difficulty levels are due to differential cognitive processes, language differences (Ercikan, 1998), solution strategies and instructional methods (Lane, Wang, Magone, 1996), and skills required by the test that are not uniformly distributed across examinees (O'Neil and McPeek, 1993).

Extensions of IRT

We have just seen how IRT extends statistical modeling beyond the constraints of classical test theory and generalizability theory. The simple elements in the basic equation of IRT (Equation 4) can be elaborated in several ways, each time expanding the range of assessment situations to which probability-based reasoning can be applied in the pursuit of psychometric principles.

Multiple-category responses. Whereas IRT was originally developed with dichotomous (right/wrong) test items, researchers have extended the machinery to observations that are coded in multiple categories. This is particularly useful for performance assessment tasks that are evaluated by raters on, say, 0-5 scales. Samejima (1969) carried out pioneering work in this regard. Thissen and Steinberg (1986) explain

the mathematics of the extension and provide a useful taxonomy of multiple-category IRT models, and Wright and Masters (1982) offer a readable introduction to their use.

Rater models. The preceding paragraph mentioned that multiple-category IRT models are useful in performance assessments with judgmental rating scales. But judges themselves are sources of uncertainty, as even knowledgeable and well-meaning raters rarely agree perfectly. Generalizability theory, discussed earlier, incorporates the overall impact of rater variation on scores. Adding terms for individual raters into the IRT framework goes further, so that we can adjust for their particular effects, offer training when it is warranted, and identify questionable ratings with greater sensitivity. Recent work along these lines is illustrated by Patz and Junker (1999) and Linacre (1989).

Conditional dependence. Standard IRT assumes that responses to different items are independent once we know the item parameters and examinee's θ. This is not strictly true when several items concern the same stimulus, as in paragraph comprehension tests. Knowledge of the content tends to improve performance on all items in the set, while misunderstandings tend to depress all, in ways that don't affect items from other sets. Ignoring these dependencies leads one to overestimate the information in the responses. The problem is more pronounced in complex tasks when responses to one subtask depend on results from an earlier subtask, or when multiple ratings of different aspects of the same performance are obtained. Wainer and his colleagues (e.g., Wainer & Keily, 1987; Bradlow, Wainer, & Wang, 1999) have studied conditional dependence in the context of IRT. This line of work is particularly important for tasks in which several aspects of the same complex performance must be evaluated (Yen, 1993).

Multiple attribute models. Standard IRT posits a single proficiency to "explain" performance on all the items in a domain. One can extend the model to situations in which multiple aspects of knowledge and skill are required in different mixes in different items. One stream of research on multivariate IRT follows the tradition of factor analysis, using analogous models and focusing on estimating structures from tests more or less as they come to the analyst from the test developers (e.g., Reckase, 1985). Another stream starts from multivariate conceptions of knowledge, and constructs tasks that contain evidence of that knowledge in theory-driven ways (e.g., Adams, Wilson, & Wang, 1997).

As such, this extension fits in neatly with the task-construction extensions discussed in the following paragraph. Either way, having a richer syntax to describe examinees within the probability-based argument supports more nuanced discussions of knowledge and the ways it is revealed in task performances.

Incorporating item features into the model. Embretson (1983) not only argued for paying greater attention to construct representation in test design, she argued for how to do it: Incorporate task model variables into the statistical model, and make explicit the ways that features of tasks impact examinees' performance. A signal article in this regard was Fischer's (1973) linear logistic test model, or LLTM. The LLTM is a simple extension of the Rasch model shown above in Equation 5, with the further requirement that each item difficulty parameter β is the sum of effects that depend on the features of that particular item:

$$\beta_j = \sum_{k=1}^{m} q_{jk}\eta_k,$$

where h_k is the contribution to item difficulty from Feature k, and q_{jk} is the extent to which Feature k is represented in Item j. Some of the substantive considerations that drive task design can thus be embedded in the statistical model, and the tools of probability-based reasoning are available to examine how well they hold up in practice (validity), how they affect measurement precision (reliability), how they can be varied while maintaining a focus on targeted knowledge (comparability), and whether some items prove hard or easy for unintended reasons (fairness). Embretson (1998) walks through a detailed example of test design, psychometric modeling, and construct validation from this point of view. Additional contributions along these lines can be found in the work of Tatsuoka (1990), Falmagne and Doignon (1988), Pirolli and Wilson (1998), and DiBello, Stout, and Roussos (1995).

Progress on other fronts

The steady extension of probability-based tools to wider ranges of assessment uses has not been limited to IRT. In this section we will mention some other important lines

of development, and point to work that is bringing these many lines of progress into the same methodological framework.

Latent class models. Research on learning suggests that knowledge and skills in some domains could be characterized as discrete states (e.g., the "production rule" models John Anderson uses in his intelligent tutoring systems--Anderson, Boyle, & Corbett, 1990). Latent class models characterize an examinee as a member of one of a number of classes, rather than as a position on a continuous scale (Lazarsfeld, 1950; Dayton, 1999; Haertel, 1989). The classes themselves can be considered ordered or unordered. The key idea is that students in different classes have different probabilities of responding in designated ways in assessment settings, depending on their values of the knowledge and skill variables that define the classes. When this is what theory suggests and purpose requires, using a latent class model offers the possibility of a more valid interpretation of assessment data. The probability-based framework of latent class modeling again enables us to rigorously test this hypothesis, and to characterize the accuracy with which observed responses identify students with classes. Reliability in latent class models is therefore expressed in terms of correct classification rates.

Models for other kinds of data. All of the machinery of IRT, including the extensions to multivariate student models, raters, and task features, can be applied to data other than just dichotomous and multiple-category observations. Less research and fewer applications appear in the literature, but the ideas can be found for counts (Rasch, 1960/1980), continuous variables (Samejima, 1973), and behavior observations such as incidence and duration (Rogosa & Ghandour, 1991).

Models that address interrelationships among variables. The developments in measurement models we have discussed encompass wider ranges of student models, observations, and task features, all increasing the fidelity of probability-based models to real-world situations. This contributes to improved construct representation. Progress on methods to study nomothetic span has taken place as well. Important examples include structural equations models and hierarchical models. Structural equations models (e.g., Jöreskog & Sörbom, 1979) incorporate theoretical relationships among variables and simultaneously take measurement error into account, so that complex hypotheses can be

posed and tested coherently. Hierarchical models (e.g., Bryk & Raudenbush, 1992) incorporate the ways that students are clustered in classrooms, classrooms within schools, and schools within higher levels of organization, to better sort out the within and across-level effects that correspond to a wide variety of instructional, organizational, and policy issues, and growth and change. Clearer specifications and coherent statistical models of the relationships among variable help researchers frame and critique "nomothetic net" validity arguments.

Progress in statistical methodology. One kind of scientific breakthrough is to recognize situations previously handled with different models, different theories, or different methods as special cases of a single approach. The previously mentioned models and accompanying computer programs for structural equations and hierarchical analyses qualify. Both have significantly advanced statistical investigations in the social sciences, validity studies among them, and made formerly esoteric analyses more widely available. Developments taking place today in statistical computing are beginning to revolutionize psychometric analysis in a similar way.

Those developments comprise resampling-based estimation, full Bayesian analysis, and modular construction of statistical models (Gelman, Carlin, Stern, and Rubin, 1995). The idea is this. The difficulty of managing evidence leads most substantive researchers to work within known and manageable families of analytic models; that is, ones with known properties, available procedures, and familiar exemplars. All of the psychometric models discussed above followed their own paths of evolution, each over the years generating its own language, its own computer programs, its own community of practitioners. Modern computing approaches such as Markov Chain Monte Carlo estimation, provide a general approach to construct and fit such models with more flexibility, and see all as variations on a common theme. In the same conceptual framework and with the same estimation approach, we can carry out probability-based reasoning with all of the models we have discussed.

Moreover, we can mix and match components of these models, and create new ones, to produce models that correspond to assessment designs motivated by theory and purpose. This approach stands in contrast to the compromises in theory and methods that

result when we have to gather data to meet the constraints of specific models and specialized computer programs. The freeware computer program BUGS (Spiegelhalter, et al., 1995) exemplifies this building-block approach. These developments are softening the boundaries between researchers who study psychometric modeling and those who address the substantive aspects of assessment. A more thoughtful integration of substantive and statistical lines of evidentiary arguments in assessment will further the understanding and the attainment of psychometric principles.

Conclusion

These are days of rapid change in assessment.[vi] Advances in cognitive psychology deepen our understanding of how students gain and use knowledge (National Research Council, 1999). Advances in technology make it possible to capture more complex performances in assessment settings, by including, for example, simulation, interactivity, collaboration, and constructed response (Bennett, 2001). Yet as forms of assessment evolve, two themes endure: The importance of psychometric principles as guarantors of social values, and their realization through sound evidentiary arguments.

We have seen that the quality of assessment depends on the quality of the evidentiary argument, and how substance, statistics, and purpose must be woven together throughout the argument. A conceptual framework such as the assessment design models of Figure 1 helps experts from different fields integrate their diverse work to achieve this end (Mislevy, Steinberg, Almond, Haertel, and Penuel, in press). Questions will persist, as to 'How do we synthesize evidence from disparate sources?', 'How much evidence do we have?', 'Does it tell us what we think it does?', and 'Are the inferences appropriate for each student?' The perspectives and the methodologies that underlie psychometric principles—validity, reliability, comparability, and fairness—provide formal tools to address these questions, in whatever specific forms they arise.

Notes

[i] We are indebted to Prof. David Schum for our understanding of evidentiary reasoning, such as it is. This first part of this section draws on Schum (1987, 1994) and Kadane & Schum (1996).

[ii] $p(X_j|\theta)$ is the probability density function for the random variable X_j, given that θ is fixed at a specified value.

[iii] Strictly speaking, CTT does not address the full distributions of true and observed scores, only means, variances, and covariances. But we want to illustrate probability-based reasoning and review CTT at the same time. Assuming normality for θ and E is the easiest way to do this, since the first two moments are sufficient for normal distributions.

[iv] In statistical terms, if the parameters are identified. Conditional independence is key, because CI relationships enable us to make multiple observations that are assumed to depend on the same unobserved variables in ways we can model. This generalizes the concept of replication that grounds reliability analysis.

[v] See Greeno, Collins, & Resnick (1996) for an overview of these three perspectives on learning and knowing, and discussion of their implications for instruction and assessment.

[vi] *Knowing what students know* (National Research Council, 2001), a report by the Committee on the Foundations of Assessment, surveys these developments.

References

Adams, R., Wilson, M.R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Anderson, J.R., Boyle, C.F., & Corbett, A.T. (1990). Cognitive modelling and intelligent tutoring. *Artificial Intelligence, 42,* 7-49.

Bennett, R. E. (2001). How the internet will help large-scale assessment reinvent itself. *Education Policy Analysis*, *9*(5).

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

Brennan, R. L. (1983). *The elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brennan, R. L. (2000/in press). An essay on the history and future of reliability from the perspective of replications. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 2000. To appear in the *Journal of Educational Measurement.*

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Bryk, A. S., & Raudenbush, S. (1992*). Hierarchical linear models: Applications and data analysis methods.* Newbury Park: Sage Publications.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. Educational and *Psychological Measurement, 20,* 37-46.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 17,* 297-334.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp.147-171). Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

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

Dayton, C. M. (1999). *Latent class scaling analysis.* Thousand Oaks, CA: Sage Publications.

Dibello, L.V., Stout, W.F., & Roussos, L.A. (1995). *Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques.* In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.

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

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

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*, 543-553.

Ercikan, K., & Julian, M. (2001, in press). Classification Accuracy of Assigning Student Performance to Proficiency Levels: Guidelines for Assessment Design. *Applied Measurement in Education.*

Falmagne, J.-C., & Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology, 41*, 1-23.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner and R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-146). New York: MacMillan.

Gulliksen, H. (1950/1987). *Theory of mental tests*. New York: John Wiley/Hillsdale, NJ; Lawrence Erlbaum.

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.

Haertel, E.H., & Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R.J. Mislevy, and I.I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. J. (1993). *Principles and selected applications of item response theory.* In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). Phoenix, AZ: American Council on Education/Oryx Press.

Hambleton, R. K. & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19-39.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenzsel procedures. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Jöreskog, K. G., and Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Kadane, J.B., & Schum, D.A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence.* New York: Wiley.

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

Kelley, T.L. (1927). *Interpretation of Educational Measurements.* New York: World Book.

Kuder, G.F., & Richardson, M.W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151-160.

Lane, W., Wang., N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational measurement: Issues and practice, 15*(4), 21-27; 31.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. R. Lazarsfeld, S. A. Star, and J. A Clausen (Eds.), *Measurement and prediction* (pp.362-412). Princeton, NJ: Princeton University Press.

Levine, M., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Linacre, J. M. (1989). *Many faceted Rasch measurement*. Doctoral Dissertation. University of Chicago.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

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. *Education Researcher*, *32*(2), 13-23.

Messick, S., Beaton, A.E., & Lord, F.M. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era. *NAEP Report 83-1.* Princeton, NJ: National Assessment for Educational Progress.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the roles of task model variables in assessment design. To appear in S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice*. Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Haertel, G., & Penuel, W. (in press). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education.* Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335-374.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (in press). Making sense of data from complex assessment. *Applied Measurement in Education.*

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

National Research Council (1999). *How people learn: Brain, mind, experience, and school.* Committee on Developments in the Science of Learning. Bransford, J. D., Brown, A. L., and Cocking, R. R. (Eds.). Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R., (Eds.). Washington, DC: National Academy Press.

O'Neil, K. A., & McPeek, W. M., (1993). In P. W. Holland, & H. Wainer (Eds*.), Differential item functioning*. (pp. 255-276). Hillsdale, NJ: Erlbaum.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 221-262). New York: American Council on Education/Macmillan.

Pirolli. P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review 105*(1), 58-82.

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

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Rogosa, D.R., & Ghandour, G.A. (1991). Statistical models for behavioral observations(with discussion). *Journal of Educational Statistics, 16,* 157-252.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*, (No. 4, Part 2).

Samejima, F. (1973). Homogeneous case of the continuous response level. *Psychometrika, 38,* 203-219.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, MD: University Press of America.

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

SEPUP (1995). *Issues, evidence, and you: Teacher's guide.* Berkeley: Lawrence Hall of Science.

Shavelson, R. J., & Webb, N. W. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271-295.

Spiegelhalter, D.J., Thomas, A., Best, N.G., & Gilks, W.R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.50.* Cambridge: MRC Biostatistics Unit.

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.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-77.

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

Traub, R. E. & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4,* 517-545.

van der Linden, W. J. (1998). Optimal test assembly. *Applied Psychological Measurement, 22*, 195-202.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Keily, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 195-201.

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

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

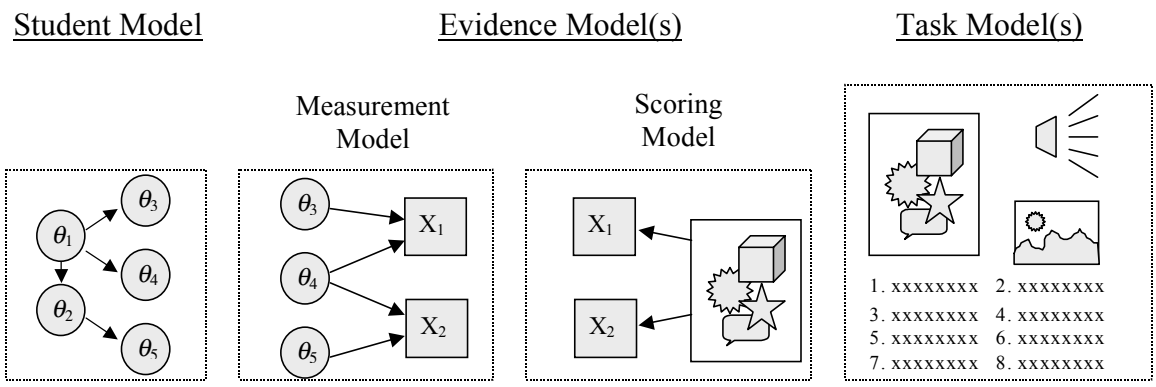

Figure 1
General Form of the Assessment Design Models

Understanding Concepts (U)--Understanding scientific concepts (such as properties and interactions of materials, energy, or thresholds) in order to apply the relevant scientific concepts to the solution of problems. This variable is the IEY version of the traditional "science content", although this content is not just "factoids".

Designing and Conducting Investigations (I)--Designing a scientific experiment, carrying through a complete scientific investigation, performing laboratory procedures to collect data, recording and organizing data, and analyzing and interpreting results of an experiment. This variable is the IEY version of the traditional "science process".

Evidence and Tradeoffs (E)--Identifying objective scientific evidence as well as evaluating the advantages and disadvantages of different possible solutions to a problem based on the available evidence.

Communicating Scientific Information (C)--Organizing and presenting results in a way that is free of technical errors and effectively communicates with the chosen audience.

Figure 2

The Variables in the Student Model for the BEAR "Issues, Evidence, and You" Example

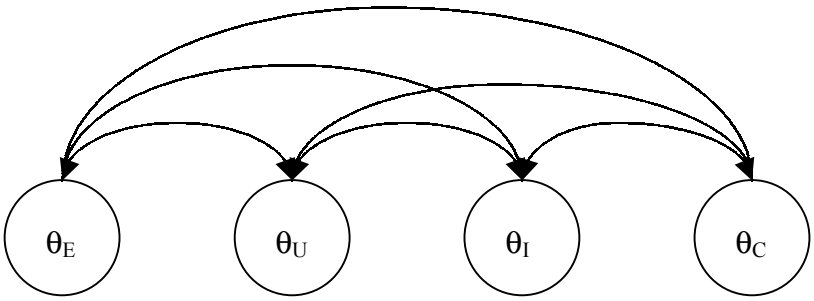

Figure 3

Graphical Representation of the BEAR Student Model

2. You are a public health official who works in the Water Department. Your supervisor has asked you to respond to the public's concern about water chlorination at the next City Council meeting. Prepare a written response explaining the issues raised in the newspaper articles. Be sure to discuss the advantages and disadvantages of chlorinating drinking water in your response, and then explain your recommendation about whether the water should be chlorinated.

Figure 4

An Example of a Task Directive from the BEAR Assessment

"As an edjucated employee of the Grizzelyville water company, I am well aware of the controversy surrounding the topic of the chlorination of our drinking water. I have read the two articals regarding the pro's and cons of chlorinated water. I have made an informed decision based on the evidence presented the articals entitled "The Peru Story" and "700 Extra People May bet Cancer in the US." It is my recommendation that our towns water be chlorin treated. The risks of infecting our citizens with a bacterial diseease such as cholera would be inevitable if we drink nontreated water. Our town should learn from the country of Peru. The artical "The Peru Story" reads thousands of inocent people die of cholera epidemic. In just months 3,500 people were killed and more infected with the diease. On the other hand if we do in fact chlorine treat our drinking water a risk is posed. An increase in bladder and rectal cancer is directly related to drinking chlorinated water. Specifically 700 more people in the US may get cancer. However, the cholera risk far outweighs the cancer risk for 2 very important reasons. Many more people will be effected by cholera where as the chance of one of our citizens getting cancer due to the water would be very minimal. Also cholera is a spreading diease where as cancer is not. If our town was infected with cholera we could pass it on to millions of others. And so, after careful consideration it is my opion that the citizens of Grizzelyville drink chlorine treated water."

Figure 5

An Example of a Student Response from the BEAR Assessment

Score	***Using Evidence:*** Response uses objective reason(s) based on relevant evidence to support choice.	***Using Evidence to Make Tradeoffs:*** Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice.
4	Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence.	Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice.
3	Response provides major objective reasons AND supports each with relevant & accurate evidence.	Response discusses <u>at least two</u> perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective.
2	Response provides <u>some</u> objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete.	Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided.
1	Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity.	Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	No response; illegible response; response offers no reasons AND no evidence to support choice made.	No response; illegible response; response lacks reasons AND offers no evidence to support decision made.
X	Student had no opportunity to respond.	

Figure 6

The Scoring Model for Evaluating Two Observable Variables from Task Responses in the BEAR Assessment

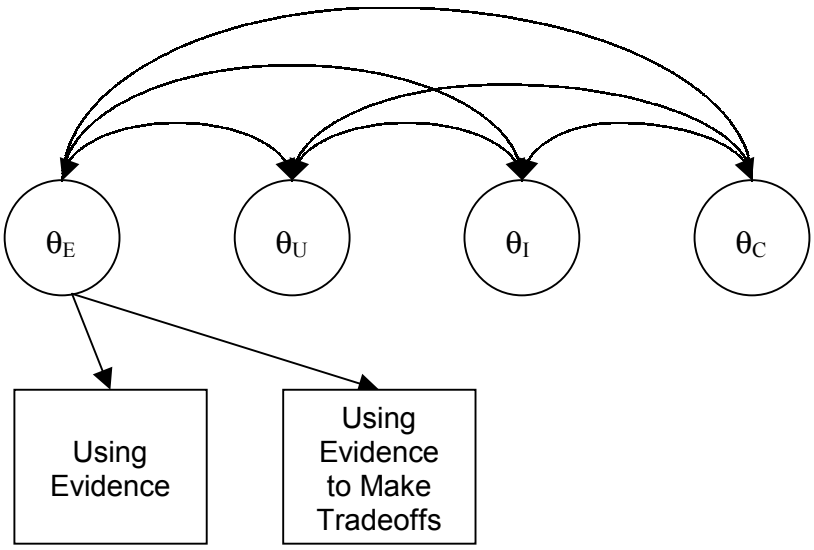

Figure 7

Graphical Representation of the Measurement Model for the BEAR Sample Task Linked to the BEAR Student Model

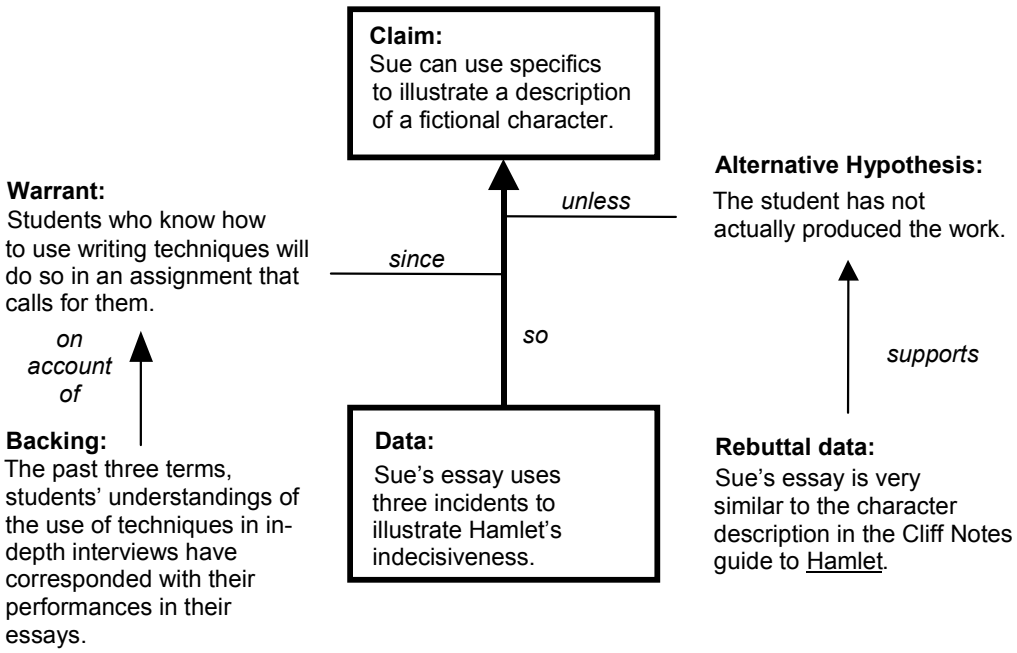

Figure 8
A Toulmin Diagram for a Simple Assessment Situation

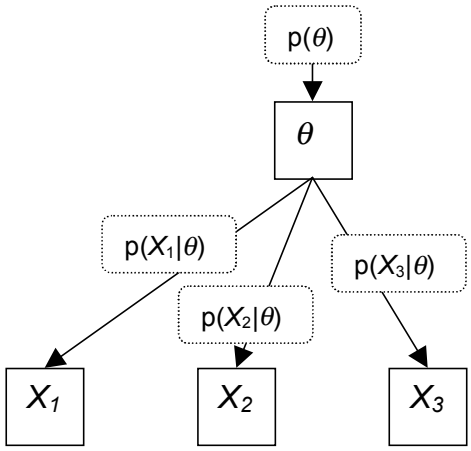

Figure 9

Statistical Representation for Classical Test Theory

Theorem

Let $N(\mu,\sigma)$ denote the normal (Gaussian) distribution with mean μ and standard deviation σ. If the prior distribution of θ is $N(\mu_0,\sigma_0)$ and the X is $N(\theta,\sigma_E)$, then the distribution for θ posterior to observing X is $N(\mu_{post},\sigma_{post})$, where $\sigma_{post}=\left(\sigma_0^{-2}+\sigma_E^{-2}\right)^{-1}$ and $\mu_{post}=\left(\sigma_0^{-2}\mu_0+\sigma_E^{-2}X\right)/\left(\sigma_0^{-2}+\sigma_E^{-2}\right)$.

Calculating the posterior distribution for Sue

Beginning with an initial distribution of $N(50,10)$, we can compute the posterior distribution for Sue's θ after seeing three independent responses by applying the theorem three times, in each case with the posterior distribution from one step becoming the prior distribution for the next step.

a) *Prior distribution*: $\theta\sim N(50,10)$.
b) *After the first response:* Given $\theta\sim N(50,10)$ and $X_1\sim N(\theta,5)$, observing X_1=70 yields the posterior $N(66.0,4.5)$.
c) *After the second response:* Given $\theta\sim N(66.0,4.5)$ and $X_2\sim N(\theta,5)$, observing X_2=75 yields the posterior $N(70.0,3.3)$.
d) *After the third response:* Given $\theta\sim N(70.0,3.3)$ and $X_3\sim N(\theta,5)$, observing X_3=85 yields the posterior $N(74.6,2.8)$.

Calculating a fit index for Sue

Suppose each of Sue's scores came from a $N(\theta,5)$ distribution. Using the posterior mean we estimated from Sue's scores, we can calculate how likely her response vector is under this measurement model using a chi-square test of fit:

$$[(70-74.6)/5]^2 + [(75-74.6)/5]^2 + [(85-74.6)/5]^2 = .85+.01+4.31 = 5.17.$$

Checking against the chi-square distribution with two degrees of freedom, we see that about 8-percent of the values are higher than this, so this vector is not that unusual.

Figure 10

A Numerical Example Using Classical Test Theory

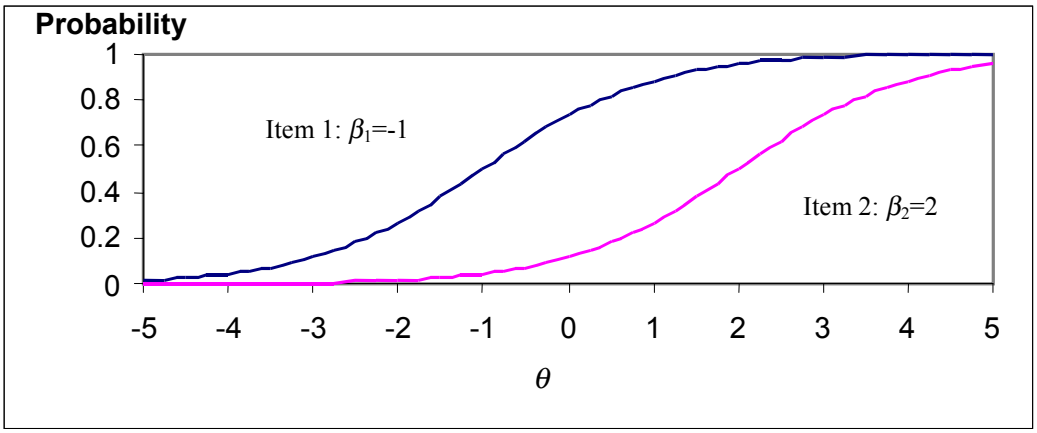

Figure 11

Two Item Response Curves

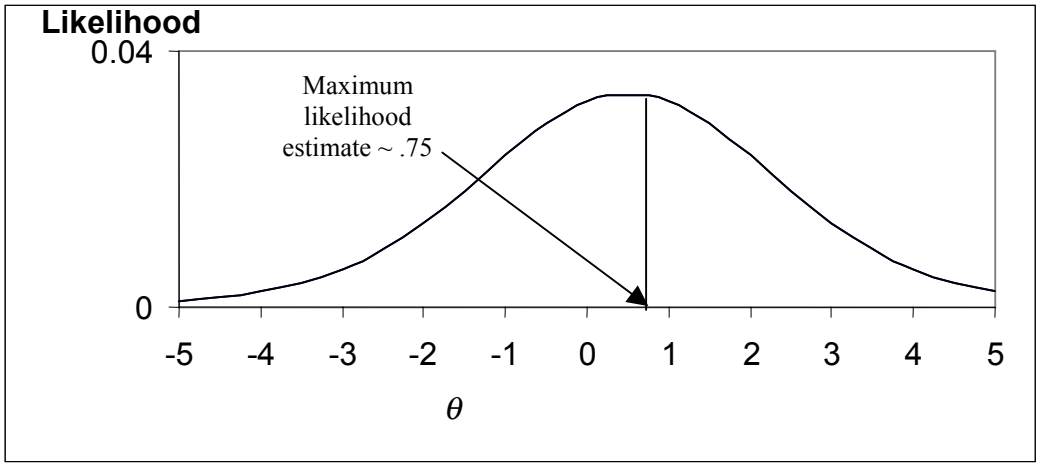

Figure 12

The IRT Likelihood Function Induced by Observing X_{i1}=0 and X_{i2}=1

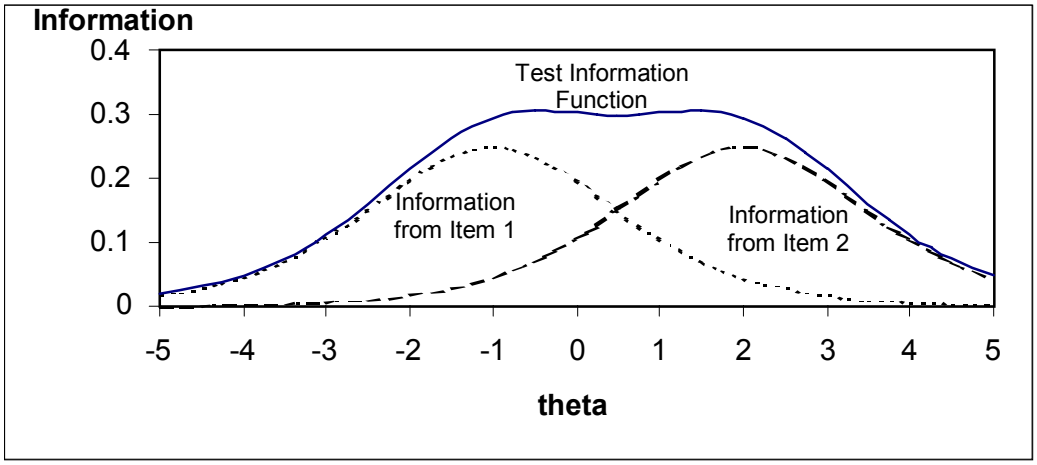

Figure 13

An IRT Test Information Curve for the Two-Item Example