Validity, 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).

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 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 quotation from Messick, above) are properties of an argument – not formulae, models, or statistics per se. We will do two things before we even introduce statistical models. First, we will look at the nature of evidentiary arguments in assessment, paying special attention to the role of standardization. And secondly, we will 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, formulae, and statistics to examine the degree to which an assessment argument possesses the salutary characteristics of validity, reliability, comparability, and fairness. Thus, it will be necessary to consider how these issues 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 on how each is used to address psychometric principles in different circumstances. We cannot provide a guidebook for using all this machinery, but we will point out some useful references along the way for the reader who needs them.

This is a long road, and it may seem to wander at times. We commence by looking at examples from an actual assessment, so that the reader will have an idea of where we want to go, in 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 the student’s knowledge or skill. Following a tradition in psychometrics, we label this “θ” (theta). This label may stand for something rather simple, such as a single category of knowledge (e.g., vocabulary usage), or something much more complex, such as a set of variables relating to the strategies that a student can bring to bear on mixed-number subtraction problems, and the conditions under which particular strategies are selected and used. The task model, at the far right in the figure 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 that we have 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.

These models may be illustrated 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