As long ago and far away as seventh-century China, standardized tests were used to inform and bolster the judgment of decision makers in selecting who should be allowed to join the imperial civil service. Even in this earliest of assessments, it was the mastery of a seemingly arbitrary skill set—the rote memorization of over 400,000 Confucian characters—that served as “evidence” of the applicant’s potential to succeed in a future job. For many centuries and all over the world, only persons of noble or otherwise elite birth had the opportunity to pursue a life of the mind or enter the halls of power, but since the academy has sought to make education a more democratically inclusive enterprise, assessments have moved toward measuring whether a person has developed those abilities presumed to have enabled prior academic success or to indicate academic aptitude. Even though there has been a shift in which developed abilities are privileged as selection criteria, the underlying process has served the same purpose: the process by which human judgment is used to make decisions concerning (1) the inclusion or exclusion of specific persons or (2) the admission or rejection of targeted persons.

When Alfred Binet was asked in 1904 to assist the government of France in distinguishing between those members of an increasingly democratic society who were educable and worthy of the nation’s investment in education, a scarce resource at that time, Binet utilized the same process. His test of intelligence was an instrument that measured the extent to which some persons had developed certain abilities, associated with success in the academy,
under the assumption that given the opportunity these persons would continue to master the declarative knowledge and procedural skills demanded of them by the institution.

In the selection of civil servants and/or the selection of persons to be admitted to the academy, the decision-making process requires human judgment. In efforts at improving those judgments, and with the rise of democracies, in efforts to reduce subjectivity and increase fairness, those involved in the assessment enterprise turned to the values that seemed to be working in the community of the scientists. As we discussed in chapter 3, the accuracy and precision on which the scientists depended came to be associated with the systematicity and standardization of measures and procedures. If the judgments and conclusions of science were to be considered valid and thus fair, the evidence—and the procedures by which the evidence was produced—needed to be considered consistent across instances and situations. Systemization and standardization were necessary to the validation of the evidence and it was the validity of the evidence that legitimatized the human decisions and judgments.

Both human judgment and assessment require reasoning from evidence. How does evidential reasoning pertain to educational assessment? To paraphrase the definitions included in the Common Core State Standards, the process of educational assessment is to synthesize data from multiple disparate sources of evidential information to make claims about the knowledge, skills, attitudes, and beliefs of students as individuals or aggregated groups. The score interpretation or assessment purpose is the claim, the data is the student behavior, and the warrant and backing are additional information about the item and the behavior of the student (e.g., existing research and theory about item performance and student cognition). It seems clear that by definition alone, an assessment is a form of an evidential argument.

Assessment is not just about measurement, it is about enabling human judgment. If what we are attempting to do in the assessment enterprise is to improve and substantiate the judgment and quality of decision making by educators and school administrators, then assessment must move beyond measurement and even beyond the qualitative analysis of measurement data (a process I will discuss in chapter 7);