PART I

THE BASIS OF COGNITIVE DIAGNOSTIC ASSESSMENT

1

Why Cognitive Diagnostic Assessment?

Jacqueline P. Leighton and Mark J. Gierl

Cognitive diagnostic assessment (CDA) is designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses. CDA is still in its infancy, but its parentage is fairly well established. In 1989, two seminal chapters in Robert Linn's *Educational Measurement* signaled both the escalating interest in and the need for cognitive diagnostic assessment. Samuel Messick's chapter, "Validity", and the late Richard Snow and David Lohman's chapter, "Implications of Cognitive Psychology for Educational Measurement", helped solidify the courtship of cognitive psychology within educational measurement. The ideas expressed in these chapters attracted many young scholars to educational measurement and persuaded other, well-established scholars to consider the potential of a relatively innovative branch of psychology, namely, cognitive psychology, for informing test development.

CDA can be traced to the ideas expressed in the previously mentioned chapters and, of course, to the many other authors whose ideas, in turn, inspired Messick, Snow, and Lohman (e.g., Cronbach, 1957; Cronbach & Meehl, 1955; Embretson, 1983; Loevinger, 1957; Pellegrino & Glaser, 1979). Since 1989, other influential articles, chapters, and books have been written specifically about CDA (see Frederiksen, Glaser, Lesgold, & Shafto, 1990). Most notably, the article by Paul Nichols (1994) titled "A Framework for Developing Cognitively Diagnostic Assessments" and the book coedited by Paul Nichols, Susan Chipman, and Robert Brennan (1995) appropriately titled *Cognitively Diagnostic Assessment*. This book, in particular, brought together a wide-ranging set of perspectives on how cognitive diagnosis might be implemented in educational measurement, including tutoring systems and job knowledge testing.

Since the mid-1980s, the marriage of *cognitive psychology*, which is focused on studying the mind in terms of the mental representations and processes that underlie observable behavior (Sternberg, 1984), and *psychometrics* has appealed to researchers and practitioners because of what it can offer. But what can it offer fundamentally? We argue that the union of cognitive psychology and psychometrics generally, and CDA in particular, offers a convincing avenue to establishing test validity. In the first section, we describe why the goal of CDA, as a way to resolve questions about test validity, is worthy of interest, investigation, and also scrutiny. To this end, we revisit some important concepts outlined by Messick (1989), Snow and Lohman (1989), and Nichols (1994) pertaining to the interest and need for unifying cognitive psychology and educational measurement. We quote, in some cases, from the original sources to illustrate and emphasize the prescience of their ideas in anticipation of CDA. We conclude this section by presenting a case for the interest and perceived need for CDA as an avenue for achieving a strong program of test validity (Cronbach & Meehl, 1955; Kane, 2001, 2006) and the information such a program offers about students' cognitive processes. In the second section, we introduce the chapters in this volume, and describe how they provide a rationale for the development of CDA and how they afford the building blocks for making such an endeavor occur in practice.

INFORMATION, TEST VALIDITY, AND COGNITIVE PSYCHOLOGY

It might be somewhat surprising that some of the most influential psychometricians in the history of testing have also been psychologists and vice versa. Robert Sternberg (1984) recounts that Sir Francis Galton was not only the inventor of the correlational method, but also an enthusiastic experimentalist in psychology, and that Alfred Binet not only created the quintessential intelligence test, but also wrote avidly about mental processes. Even Charles Spearman, who originated factor analysis, theorized intensely about cognitive processes, writing in 1923 about the "principles of cognition." And, of course, Lee Cronbach developed the most widely used measure of test reliability, while also advocating for the match between learning environments and student abilities. That some of the most prominent psychometricians in history have also been psychologists is surprising only because one would think that such complementary disciplines would be the norm today rather than the exception. However, this is not the case. Anastasi (1967) warned that "those psychologists specializing in psychometrics have been devoting more and more of their efforts to refining techniques of test construction, while losing sight of the behavior they set out to measure" (p. 297). And R. J. Mislevy (1993) added 25 years later "it is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology" (p. 19). Perhaps, as Sternberg (1984) describes, the impediment has been more fundamentally "a sociological one and resides primarily (but not exclusively) in the professional identification of the investigator and of the methods he or she uses" (p. 41). In other words, psychometricians at the end of the day must focus on *metrics* and not on *psychology*.

We are now at a moment in time, however, when even sociological impediments must be overcome. There is increasing pressure to make assessments more informative about the mental processes they measure in students. In particular, there is increasing pressure to adapt costly large-scale assessments (Organisation for Economic Co-operation and Development [OECD], 2004; U.S. Department of Education, 2004) to be informative about students' cognitive strengths and weaknesses. In the United States, for example, the No Child Left Behind Act of 2001 has made completing high-stakes, large-scale state assessments a rite of passage for almost all students and teachers. These tests are intended not only to determine students' learning outcomes and needs, but also to evaluate instructional programs (school effectiveness). In other parts of the world, the zest for educational accountability and standards to ensure that students are prepared and competitive for knowledge-based work environments has also shepherded an appetite for informative large-scale testing (OECD, 2004). What are these large-scale tests supposed to inform stakeholders of? The information being sought from these tests is essentially about students' cognitive strengths and weaknesses in thinking and learning. That is, to what extent do test scores reflect certain forms of thinking and higher-order cognitive processes associated with meaningful learning, as opposed to misconceptions and localized testwise strategies associated with lower-level understanding? Large-scale assessments are increasingly scrutinized about what they can deliver for pinpointing why students perform as they do and how students' opportunities to learn can be maximized.

Wanting to Make Inferences about Psychological Process: A Substantive Approach

In 1989, Messick already anticipated the importance of providing information about students' mental processes (as opposed to simply content-based behaviors) from test scores:

> Thus, the heart of the notion of so-called content validity is that the test items are samples of a behavioral domain or item universe about which inferences are to be drawn. But these inferences are likely to invoke, even if only tacitly, psychological processes or behaviors rather than mere surface content. (p. 36)

With these words, Messick openly expressed a reasonable observation – that the real, but unspoken, targets of inference in which many educators are interested are about students' psychological or mental processes. And why should this not be the case? Many problem-solving behaviors are known to be related to, and in some cases a direct consequence of, cognitive antecedents such as insufficient knowledge or unsophisticated strategy selection (Newell & Simon, 1972). It is sensible to want to draw test-based inferences about students' mental processes if only to increase the likelihood of providing the most effective and timely instruction to students in a way that cuts to the origin of behavior. With such information, teachers could alter students' misconceptions and replace faulty strategies. According to Messick (1989), understanding test performances substantively in terms of the mental processes students use to answer and/or solve test items is a core feature of construct validity theory. In particular, he regarded the substantive approach in construct theory as having a definitive role in the domain specification of a test:

> In the substantive approach, items are included in the original pool on the basis of judged relevance to a broadly defined domain but are selected for the test on the basis of empirical response consistencies. The substantive component of construct validity is the ability of the construct theory to account for the resultant test content. . . . the internal structure and substance of the test can be addressed more directly by means of causal modeling of item or task performance. This approach to *construct representation* attempts to identify the theoretical mechanisms that underlie task performance, primarily by decomposing the task into requisite component processes (Embretson, 1983). Being firmly grounded in the cognitive psychology of information processing, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge (including self-knowledge) that are implicated in test performance. (pp. 42–45)

However, grounding test-based inferences in the cognitive psychology of information processing is not straightforward. There is a catch, and the catch involves developing and pursuing a fairly rigorous program of construct validation (Cronbach, 1988; Kane, 2001; Messick, 1989). The test developer must begin with a well-grounded construct theory from which items will be generated and selected, and from which predictions about score relationships will be made. In considering data and analyses relevant to such a program of construct validation, Messick suggested including (a) judgmental and logical analyses to discover alternative hypotheses of score interpretation, (b) correlational or covariance analyses to search for convergent and discriminant evidence to the patterns expected from construct theory, (c) analyses of process to probe construct representation, (d) analyses of group differences and changes over time, (e) responsiveness of scores to experimental treatment and manipulation of conditions, (f) generalizability of score interpretation across contexts, and (g) threats to the tenability and generalizability of research conclusions. Notice that not only do these steps pertain to CDA, but they also directly seek to identify how students of different ability or achievement levels mentally represent and manipulate test information over time, in differing contexts, and in response to instructional interventions and test variable manipulations. However, these steps also require a committed effort of time and resources from test developers to understanding the psychology of test taking. This is the catch.

A commitment to these seven steps requires a radical shift in how testing is viewed and developed. It requires that we consider testing as a concrete, scientific endeavor instead of a circumstantial enterprise, where it is largely correlational evidence that is collected in ad hoc fashion (often after the test has been administered) to justify the interpretation of test scores (Borsboom, Mellenbergh, & Van Heerden, 2004). A quote used by Messick, which also bears repeating here, was offered by Peak (1953) more than 50 years ago: "a protest must be entered . . . against the proliferation of blindly empirical validities which are without the disciplined guidance of theory, for the increment of meaning from the accumulation of miscellaneous correlations may ultimately approach zero" (p. 288). In sum, CDA requires us to pursue a rigorous program of validation, one that is focused on measuring students' mental processes as they engage in test-taking behaviors and then using this information for improving students' opportunity to learn.

Exploring the Substantive Approach with Cognitive Psychology

Even more so than Messick, Snow and Lohman (1989) were explicit in their statements of how cognitive psychology could be used to inform educational measurement:

> *First*, the cognitive psychology of problem solving is a central concern for educational measurement because all mental tests are, in some sense, problem-solving tasks. Hence, existing or proposed test designs ought to be evaluated as such. . . . *Second*, the two most general purposes of educational measurement, the assessment of student aptitudes and achievements, would appear to cut across the matrix of cognitive psychology in different ways. . . . Thus, different slices across the field of cognitive psychology might be needed to inform test design and evaluation . . . (p. 265)

Snow and Lohman indicated that the ideas, theories, and methods of cognitive psychology could contribute to the advancement of educational measurement by (a) informing analyses of existing tests to elucidate their underlying constructs; (b) clarifying the goals of testing in terms of the knowledge and skills that are genuine indicators of mastery and understanding; and (c) enhancing theories of aptitude, achievement, and learning across different domains.

In documenting the ways in which cognitive psychology could be useful, Snow and Lohman (1989, p. 267) also recognized the important distinction between an investigator's conceptualization of a person's reasoning and problem solving versus the actual reasoning and problem solving used by the individual when responding to test items. This is a subtle but essential distinction. Such a distinction must be acknowledged in order to fully integrate the psychology of cognition in measurement. Cognitive psychologists, at least in principle, acknowledge that the most sophisticated computer models of what is expected of individual cognitive functioning must be verified with experimental studies on how individuals actually think and reason (e.g., Ericsson & Simon, 1993). Computerized models are, at best, complex hypotheses of how humans are expected to reason in the face of specific constraints. However, for computer models to truthfully inform us of the nature of human reasoning and problem solving, they must approximate real-life human thinking. This is a perspective that is not often taken into account in traditional educational psychometric models such as those embodied in classical test theory and item response theory, although it is directly applicable to developments of CDAs (Snow & Lohman, 1989; see also Embretson & Gorin, 2001; Frederiksen, Glazer, Lesgold, and Shafto, 1990; Irvine & Kyllonen, 2002; Leighton & Gierl, in press; Mislevy, Steinberg, & Almond, 2003; Nichols, 1994).

Adapting Educational Psychometric Measurement Models for Psychological Theory: Structural Fidelity

Although there is some variety in the particular assumptions made across different educational psychometric measurement (EPM) models, as Snow and Lohman refer to them, in general they aim to approximate a person's location on an underlying variable of interest such as science achievement or spatial aptitude. The location at which the person is finally placed is often interpreted as reflecting the sum or amount of the variable that that person has acquired, such as 67% of science achievement or 85% of spatial aptitude. EPM models such as those based on item response theory have contributed greatly to educational and psychological measurement by overcoming important technical obstacles (e.g., an examinee's ability estimate being dependent on the particular sample of test items chosen). However, this groundbreaking measurement has exhibited limitations in the face of changing educational contexts and climates, in which there is ever-increasing demand for information about students' cognitive processing. Serious limitations with EPM models were identified by Snow and Lohman specifically in relation to their failure to incorporate (a) substantive psychological theory to explain item responses; (b) realistic assumptions about the psychological dependencies and variables influencing test item performance (e.g., Lord's [1980] three-parameter logistic model and the assumption that only three parameters influence student item responses); and (c) explicit delineation of the psychological processes that collectively reflect the construct measured by a test. In addition, the implicit cognitive models that inform many educational tests are still reflective of investigators' *expectations* of how students will reason and solve problems in test-taking situations; they are not based on empirical evidence of how students actually think in these circumstances (Leighton & Gierl, in press; Nichols, 1994).

The limitations that EPM models exhibit must be addressed and ultimately overcome for the successful use of CDA. Of course, the most sophisticated of substantive theories would likely be of little use to the development of CDA if such theories could not be incorporated into psychometric models. However, EPM models must now be *adapted* to assimilate and accommodate substantive components of test-taking behavior. Messick (1989) described the value of developing the proper EPM models:

> Subsequently, Loevinger (1957) formalized the call for rational scoring models by coining the term *structural fidelity*, which refers to "the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured" (p. 661). The structural component of construct validity includes both this fidelity of the scoring model to the structural characteristics of the construct's nontest manifestations and the degree of interitem structure. (p. 43)

It is tempting to imagine that cognitive psychology can be infused in EPM models directly and, with minor but clever tweaking, transform EPM models completely. Unfortunately, it is not quite that simple. Cognitive theories exist for many phenomena, including perception, memory, attention, reasoning, problem solving, intelligence, and even special abilities (Healy, 2005). However, there are few, if any, cognitive theories about assessment in particular; or about the multifaceted and complex test-taking processes and behaviors that educational tests aim to measure; or about achievement generally as measured by educational assessments. What this means is that the challenge Snow and Lohman (1989) saw for cognitive psychology to "develop improved substantive theories of the aptitudes and achievements that are the goals of education and that educational measurements should be designed to assess and promote" (p. 269) is still true today. That is, substantive theories and empirical studies are still in demand. Borrowing theories from cognitive psychology and importing them into educational measurement initiatives is possible but difficult because these theories are largely developed within narrow learning contexts and in the absence of formal assessment frameworks (e.g., John R. Anderson's ACT programming tutor; see Anderson, Corbett, Koedinger, & Pelletier, 1995, and Anderson & Gluck, 2001). It would be ideal if cognitive psychologists were to develop such theories solely for educational measurement tasks or at least with educational measurement in mind. But we think the reality is largely the same as Sternberg (1984) described it – scholars identify quite narrowly with their own domains of interest and methods. Consequently, the onus is on educational researchers to adapt our methods, techniques, and tools to incorporate cognitive theory and, more important, to actively create, modify, and test theories of cognition for educational measurement purposes. In other words, *we* need to put educational tests under the "cognitive microscope."

A Conceptual Recipe for Developing Cognitive Diagnostic Assessments

The cognitive microscope magnifies the substantive component of educational assessments (Messick, 1989; Snow & Lohman, 1989). This focal point forces us to explicitly consider the construct representation and the structural fidelity of the assessment against the learning process we aim to measure. Putting educational assessments under the cognitive microscope emphasizes the appeal of CDAs as Nichols (1994) articulated so well:

> These new assessments [cognitive diagnostic assessments] make explicit the test developer's substantive assumptions regarding the processes and knowledge structures a performer in the test domain would use, how the processes and knowledge structures develop, and how more competent performers differ from less competent performers. (p. 578)

In other words, educational tests designed for cognitive diagnostic purposes are different from traditional approaches in that they do not rely solely on logical taxonomies and content specifications to describe their objectives. This is because "efforts to represent content are only vaguely directed at revealing mechanisms test takers use in responding to items or tasks" (p. 585). Instead, educational tests designed for cognitive diagnosis rely largely on, and are informed by, the psychology of learning, reasoning, and problem solving to describe their purpose. To this end, Nichols outlined five steps for psychology-driven test development:

1. *Substantive theory construction*: This first step requires the development of a model or theory that characterizes the hypothesized knowledge structures and processes required to perform (respond to) the assessment. In addition, the item variables that invoke particular cognitive processes and knowledge structures must be identified.
2. *Design selection*: This second step, guided by the model or theory developed in step 1, requires the test developer to choose the observation and measurement design. The test items chosen will be selected (or created) with the expectation that test takers will respond in predictable ways, with the processes and knowledge structures identified in step 1, to the items.
3. *Test administration*: This third step involves important details of the environment and context in which test takers complete their assessments such as item format, medium of item presentation,