Defining and Evaluating Models of Cognition Used in Educational Measurement to Make Inferences About Examinees' Thinking Processes

Jacqueline P. Leighton and Mark J. Gierl, *University of Alberta*

The purpose of this paper is to define and evaluate the categories of cognitive models underlying at least three types of educational tests. We argue that while all educational tests may be based—explicitly or implicitly—on a cognitive model, the categories of cognitive models underlying tests often range in their development and in the psychological evidence gathered to support their value. For researchers and practitioners, awareness of different cognitive models may facilitate the evaluation of educational measures for the purpose of generating diagnostic inferences, especially about examinees' thinking processes, including misconceptions, strengths, and/or abilities. We think a discussion of the types of cognitive models underlying educational measures is useful not only for taxonomic ends, but also for becoming increasingly aware of evidentiary claims in educational assessment and for promoting the explicit identification of cognitive models in test development. We begin our discussion by defining the term cognitive model in educational measurement. Next, we review and evaluate three categories of cognitive models that have been identified for educational testing purposes using examples from the literature. Finally, we highlight the practical implications of "blending" models for the purpose of improving educational measures.

Keywords: cognitive models, construct validity, diagnostic testing, educational measurement

It would not be an exaggeration to say that the new millennium has brought an unprecedented demand for testing at all levels of the educational system. In the United States alone, the No Child Left Behind (NCLB) Act of 2001, considered to be the most extensive reform of the Elementary and Secondary Education Act (ESEA) of 1965, has come to dominate discussions and administrations of assessment policies and products. The NCLB Act has effectively redefined the federal government's role in kindergarten through grade 12 education. The depth and breadth of information that large-scale educational tests are now expected to provide to stakeholders is staggering. In describing the NCLB Act, the United States Department of Education claims:

> The new law will empower parents, citizens, educators, administrators and policymakers with data from those annual assessments. . .The tests will give teachers and principals information about how each child is performing and help them to diagnose and meet the needs of each student. (USDE, 2004, 3rd paragraph)

Extraordinary claims about the informational value of large-scale educational tests are increasingly being endorsed outside the United States as well. For example, the Organization for Economic Co-operation and Development (OECD) launched, in 1997, the Programme for International Student Assessment (PISA) with the intent of measuring how well 15-year-old students were prepared to meet the challenges of work and innovation in knowledge-based economies:

> The assessment is forward-looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum. (OECD, 2004, p. 20)

The objective of this paper is to examine different kinds of educational tests and the psychological evidence they have for claims (e.g., inferences and diagnoses) of examinees' thinking processes in relation to their learning, achievement, and general academic abilities.[1] We will focus on three types of tests—large-scale, classroom-based,

Jacqueline P. Leighton and Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada T6G 2G5; jacqueline.leighton@ualberta.ca.

and cognitive diagnostic—along with the corresponding categories of cognitive models that normally underwrite their test-based inferences.[2] By examining these three types of educational tests, we intend to illustrate the challenges with using results from large-scale tests, normally in the form of international, national, state, and district tests, to deliver diagnostic information about examinees' thinking processes for learning (Leighton & Gierl, in press-a). We begin our discussion by commenting on how the term *cognitive model* should be defined in educational measurement. Next, we review and critically evaluate three categories of cognitive models that have been identified for informing three types of educational tests, using examples from the literature. These three cognitive models differ in the specificity of the thinking processes they represent and in the strength of psychological evidence they provide to support diagnostic claims about examinees' cognitive strengths and weaknesses. Finally, we highlight the type of cognitive model that might be able to yield strong diagnostic inferences about examinees' thinking processes and present some of the challenges with blending this model into large-scale testing programs.

Prologue and Disclaimer

We concur with the National Research Council (2001) that cognitive models of learning are an essential part of the framework used to characterize the construct of a test, design test items, and generate inferences about examinees from their test performance (see also Kane, 1992; Messick, 1989; Mislevy, Steinberg, & Almond, 2003). In an ideal world, test developers would adhere explicitly to the steps outlined by the National Research Council (NRC, 2001) and ensure that "every assessment is grounded in a conception or theory about how people learn, what they know, and how knowledge and understanding progress over time" (p. 20). In this ideal world, few would likely argue with the development and use of existing large-scale tests for generating inferences about examinees' thinking processes, and the NRC and other similar agencies would find it unnecessary to suggest that:

> On the whole, most current large-scale tests provide very limited information that teachers and educational administrators can use to identify why students do not perform well or to modify the conditions of instructions in ways likely to improve student achievement. (NRC, 2001, p, 27)

It is our supposition that if diagnostic claims about examinees' cognitive strengths and weaknesses are assumed to be justifiable from current large-scale tests (as appears to be the case from the statements made by the USDE and OECD), then some ambiguity must exist as to the kinds of tests (and cognitive models) that do in fact support diagnostic claims of students' thinking and/or learning processes.

We think that by contrasting at least three types of tests and the cognitive models that inform them, a clearer understanding can be achieved about the nature of the cognitive models underlying these tests and the inferences they can defensibly support. Announcements suggesting the adequacy of large-scale tests for "[helping teachers and principals] to diagnose and meet the needs of each student" (USDE, 2004, 3rd paragraph) or providing information about "young people's ability to use their knowledge and skills to meet real-life challenges" (OECD, 2004, p. 18) imply that these large-scale tests can be used to generate strong claims about examinees' thinking processes in academic domains. We are aware of no research endorsing the use of results from large-scale tests to support strong claims about examinees' thinking processes. Consequently, advertising or even insinuating such claims may do a substantial disservice to the public about what large-scale tests can and *cannot* do. As such, clarifying the nature of the categories of cognitive models underlying tests in order to scrutinize whether they can support claims about examinees' thinking processes is a meaningful endeavor in so far as it explores the evidential basis of tests and the validity of inferences. In our illustration of the cognitive models underlying three types of tests, we will not immediately argue for whether the cognitive models should be blended in the design of future assessments. This should not be viewed as our acceptance of the status quo but, rather, as an attempt to first clarify the murky state of affairs in the definition and current use of cognitive models in educational measurement. Although the blending of cognitive models may seem to be a natural step in the improvement of assessments, we will deal with the challenges facing this aspect of test development in the Implications section of the paper.

Defining Cognitive Models in Educational Measurement

The Demand for Information about Examinees' Thinking Processes

Traditionally only classroom-based tests have been used for formative purposes, helping teachers monitor student learning and achievement. Since the NCLB Act, however, classroom-based tests are no longer expected to be the sole source of this kind of information. If the quotes illustrated previously from the U.S. Department of Education and the OECD are taken at face value, then large-scale tests have been anointed with a similar purpose (see also with respect to second language learning and the SAT/GRE: Buck, Tatsuoka, & Kostin, 1997; Embretson & Wetzel, 1987; Gierl, Tan, & Wang, 2005; Gorin, 2005; Luecht, 2005; Thissen & Edwards, 2005; Wang, Deng, Williams, & Laitusis, 2005). Indeed, there is growing demand for using large-scale achievement test results to generate better, more specific inferences about examinees (e.g., Embretson, 1998, 1999; Embretson & Gorin, 2001; Haladyna & Downing, 2004; Irvine & Kyllonen, 2002; Mislevy, 1996; Pearson & Garavaglia, 2003; Pellegrino, Baxter, & Glaser, 1999; Sternberg, 1984)—not simply for a rank, but for specifying what has become somewhat of a cliché—identifying examinees' cognitive strengths and weaknesses. However, few large-scale tests are able to yield diagnostic information about examinees' thinking processes (including their strengths and weaknesses) because few large-scale tests are developed with these explicit targets of inference (Lane, 2004; Leighton, 2004; Leighton, Gierl, & Hunka, 2004; see also Mislevy et al., 2003; Nichols, Chipman, & Brennan, 1995; Norris, Leighton, & Phillips, 2004; NRC, 2001; Snow, 1993). To support diagnostic inferences about examinees' thinking processes, large-scale tests must be developed from empirically-based cognitive models of learning (Nichols et al., 1995; Norris et al., 2004; NRC, 2001; Snow & Lohman, 1989).

Although the use of cognitive models may be considered necessary for the development of information-rich educational tests, the use of cognitive models remains limited by at least two

issues. First, what constitutes a *cognitive model* for educational measurement remains unclear. On the one hand, a cognitive model could be assumed to be any set of beliefs about student learning even in the absence of supporting scientific evidence. A set of beliefs could be considered a cognitive model simply because it is an informal, but plausible, theory about how students *might* think. We all have such beliefs; they correspond to what psychologists call our *theories of mind* (Wellman & Lagattuta, 2004). Translating this set of beliefs into a cognitive model is common in educational measurement. For example, the knowledge content experts provide to testing companies about the conceptual parameters of learning domains and whether test items reflect learning objectives is highly valuable in the development of tests. In fact, the knowledge and experience these content experts possess is sufficiently seductive as to be often extended beyond simply informing the conceptual parameters of learning domains, and toward speculating how students will think about academic tasks, including the skills examinees will use to solve educational test items (Norris et al., 2004). Such notions of how students will think about and solve academic tasks may suffice to generate a simple cognitive model of student learning. However, such notions, albeit sophisticated theories of mind, remain unsubstantiated unless they are tested empirically. On the other hand, a cognitive model could be assumed to be any theory that has extensive scientific backing achieved through widespread psychological investigations. It is currently not clear what should be the extent of scientific backing for cognitive models underlying educational tests.

Second, assuming these definitional issues can be resolved, the characteristics of cognitive models require evaluation so as to avoid conflating the different categories of cognitive models with the diagnostic claims (about students) they can defensibly support. In particular, it seems necessary to evaluate whether large-scale tests are currently developed from cognitive models that can underwrite claims to help "diagnose and meet the needs of each student" (USDE, 2004, 3rd paragraph) or provide information about "young people's ability to use their knowledge and skills to meet real-life challenges" (OECD, 2004, p. 18). Although these claims advertise the desperate and increasing demand for higher-level inferences about examinees from their large-scale test results, we need to take a temperate look at whether these tests are developed from cognitive models that can indeed deliver such news.

A Definition of Cognitive Model

Students' thinking on academic tasks cannot be observed and evaluated directly. Rather, students' thinking must be judged indirectly from their performance on a given task. From a student's correct answer on an educational task, for example, we make the assumption that the student engaged in a specific and correct sequence of thought to generate the answer. We make this assumption to justify predictions about the student in future situations, as we predict that the student will reproduce this correct sequence of thought and be successful on future tasks of similar content and complexity (Norris et al., 2004). This is a convenient assumption, but it is unsubstantiated in many cases (Cronbach, 1971; Leighton & Gokiert, 2005; Norris et al., 2004; Poggio et al., 2005; Rogers & Harley, 1999; Rogers & Yang, 1996; Snow, 1993). What many studies are beginning to show is that students can generate correct answers using patterns of thought that are unrelated to the knowledge and skills targeted by the test item (e.g., Gierl, 1997; Leighton & Gokiert, 2005; Poggio et al., 2005). Genuine domain mastery or competence can be usurped by test-wise strategies and alternate knowledge not specifically targeted by the test item (Rogers & Yang, 1996). Moreover, these alternate patterns of thought may not generalize to successful performance in future situations on similar tasks.

Most, if not all, measures of academic achievement are assumed to be based on some category of cognitive model. The National Research Council (NRC) in 2001, in conjunction with an esteemed group of educational researchers, emphasized this when it stated that:

> every assessment, regardless of its purpose, rests on three pillars: a model of how students represent knowledge and develop competence in the subject domain [a cognitive model of learning], tasks or situations that allow one to observe students' performance, and an interpretation method for drawing inferences from the performance evidence thus obtained. (p. 2)

The NRC further characterized a cognitive model of learning as one that is (a) based on empirical studies of learning in the domain, (b) able to distinguish between beginning and skilled learners, (c) formulated from the variety of ways in which students develop understanding of the subject matter, and (d) amenable to aggregation in a principled manner so that it can be used for different assessment purposes (pp. 178–185). This description of a cognitive model seems clear, but if we accept this description, we are forced to recognize a contradiction—namely, that not all assessments are based on cognitive models of learning. In particular, large-scale tests are not based on cognitive models because they are not developed from empirical studies of student learning. So if we accept the NRC's proposal that *all* assessments do indeed rest on three pillars, one of which is a cognitive model of learning, then we must consider a less restrictive definition of a cognitive model of learning so as to include the large majority of large-scale tests.

The term *cognitive model* originates in the field of computer science where it is defined as the simulation of human problem solving and mental task processing. The term is defined similarly within cognitive psychology as a simplified description of human problem solving often assuming a computational model that is substantiated with human studies of cognitive processing (e.g., Anderson, Qin, Sohn, Stenger, & Carter, 2003; Baddeley & Logie, 1999; Ericsson & Simon, 1993; Healy, 2005; Kalchman, Moss, & Case, 2001; Newell & Simon, 1972; Siegler, 2005). Cognitive models have been useful in predicting and explaining information-processing procedures for a variety of problem-solving behaviors (e.g., Ericsson & Simon, 1993). Because the cognitive psychological definition often assumes a computational model that is substantiated with human studies of cognitive processing, it poses a strict standard (similar to the description by the NRC) and, therefore, a potential usability problem for educational measurement researchers and practitioners. It is understandable that this standard should exist for cognitive psychologists because they are motivated to make specific inferences about underlying cognitive structures and response processes—even at the cost of focusing on a very narrow range of knowledge

and skills. However, for educational measurement researchers and practitioners this standard may be too restrictive, especially because test-based inferences are typically broad, covering multiple knowledge and skill domains.

If we are to use the term cognitive model in educational measurement, then it may be necessary to modify, slightly, the cognitive psychological definition so that it also pertains to the testing of broad problem-solving knowledge and skills. Thus, we suggest the term cognitive model in educational measurement refer to a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance." This definition does not include any mention of computational models or human studies because, if it did, it could not be extended to encompass most current large-scale tests in educational measurement. However, by leaving out the requirements of computational models or human studies in our definition, it is possible to stay true to the NRC's indication that "every assessment is grounded in a conception or theory about how people learn, what they know, and how knowledge and understanding progress over time" (p. 20). Moreover, our definition does not include explicit mention of students' communicative practices because it is assumed that a cognitive model developed for a population of interest should and will consider the communicative processes of that population. Now we turn to the question of which cognitive models do indeed lend support to diagnostic inferences about students' thinking processes.

Evaluating Three Cognitive Models in Educational Measurement

Elaborating on the NRC's recommendation to identify cognitive models of learning in educational measurement, Leighton (2004) identified three categories of models that could be used to describe educational achievement tests. In this section, we elaborate on these three models with illustrative examples from the educational measurement literature. Because there are costs and benefits associated with each of the three models, we critically evaluate each model in terms of the specificity of the knowledge and skills covered, and the psychological evidence for supporting diagnostic inferences about examinees' thinking processes. The purpose of this section is not to sanction the use of one model over another. Rather, our interest is to clarify the categories of cognitive models underlying different types of tests and, more importantly, specify the type of inferences that can be defensibly made. As mentioned previously, the identification of the cognitive models underlying educational tests is helpful because it encourages the test developer and practitioner to consider the alignment between the model and the test against the target of inference. Depending on the target of inference, the type of cognitive model developed or used for a test will need to change. Not all cognitive models will support similar types of inferences.

The Cognitive Model of Test Specifications

Description

The first cognitive model is of *test specifications.* Researchers and practitioners in educational measurement will be familiar with this model given that it is often used to develop tests. In its most common form, a model of test specifications is generated using a two-way matrix to establish content and skill groupings for obtaining a *representative sample* of items during test construction from a defined achievement domain. In this two-way matrix, the rows might represent content areas and the columns might represent skills to be measured by the test. Test items are then generated to represent each combination of content and skill in the matrix, anticipating the cognitive skills examinees will use in each content area (Bloom, 1956; Gierl, 1997; Webb, 2006). Large-scale standardized tests are often representative of the cognitive model of test specifications. Although different varieties of large-scale tests exist, these assessments are often designed by professional test developers in government or testing companies to evaluate examinees' general skill attainments at specified grade levels.

An illustrative example of a large-scale test developed from a model of test specifications can be found in the subject tests administered by the College Entrance Examination Board. These subject tests, formerly known as the SAT II: Subject Tests, are designed to measure examinees' core knowledge and skills in particular subject areas, as well as their ability to apply that knowledge. The tests are independent of any particular textbook or method of instruction (College Entrance Examination Board, 2006). The content of the tests evolves to reflect current trends in high school curricula, but the types of questions change little from year to year. The sequence of steps followed to develop the SAT Subject test is routine, involving an appointed committee of secondary school teachers and college faculty to help design and scrutinize test items. The content and skill specifications for the subject test in Physics are shown in Figure 1. Content experts (or subject-matter specialists) and statistical specialists help the development committee refine the test by providing technical feedback on test items, field-testing, and data analysis.

Evaluation

The benefit of the cognitive model of test specifications is its simplicity because committees of experts can generate it and approve it. Human studies are rarely conducted to verify that the thinking processes examinees apply to the test are well aligned to the expectations outlined in the test specifications. Although human studies are not often conducted, the test specifications nonetheless function as a cognitive model because they not only reflect an explicit demarcation of the domain of achievement, but also, more importantly, the knowledge and skills examinees *are expected to use to answer test items correctly*. Although viewing test specifications as a type of cognitive model may initially seem overly permissive, possibly opening the door to almost anyone claiming that they have adhered to the NRC recommendations, we believe this is not the case. Any cognitive model is simply a working hypothesis of how one group of people (e.g., test developers or cognitive scientists) believes another group of people (e.g., students or research participants) is thinking and thus solving a set of test items or tasks. The model of test specifications is therefore nothing more than a working hypothesis of students' knowledge and skills at a general level of detail. In fact, the generality of the model of test specifications is often considered economical and beneficial. The broad sampling of behaviors, coupled with its

Content: Physics		
Mechanics	Approx. % of Test	Approx. No. of Items
• Kinematics (e.g., velocity, acceleration, motion in one dimension, and motion of projectiles) • Dynamics (e.g., force, Newton's laws, and statics) • Energy and Momentum (e.g., potential and kinetic energy, work, power, impulse, and conservation laws) • Circular Motion (e.g., uniform circular motion, centripetal force) • Simple Harmonic Motion (e.g., mass on a spring and the pendulum) • Gravity (e.g., the law of gravitation, orbits, and Kepler's laws)	36-42	27-32
Electricity and Magnetism		
• Electric Fields, Forces, and Potentials (e.g., Coulomb's law, induced charge, field and potential of groups of point charges, and charged particles in electric fields) • Capacitance (e.g., parallel-plate capacitors and transients) • Circuit Elements and DC Circuits (e.g., resistors, light bulbs, series and parallel networks, Ohm's law, Joule's law) • Magnetism (e.g., permanent magnets, fields caused by currents, particles in magnetic fields, Faraday's law, and Lenz's law)	18-24	14-18
Waves		
• General Wave Properties (e.g., wave speed, frequency, wavelength, superposition, standing waves, and Doppler effect) • Reflection and Refraction (e.g., Snell's law and changes in wavelength and speed) • Ray Optics (e.g., image formation using pinholes, mirrors, and lenses) • Physical Optics (e.g., single-fit diffraction, double-sit interference, polarization, and color)	15-19	11-14
Heat and Thermodynamics		
• Thermal Properties (e.g., temperature, specific and latent heats, thermal expansion, and heat transfer) • Law of Thermodynamics (e.g., first and second laws, internal energy, and heat engine efficiency)	6-11	5-8
Modern Physics		
• Quantum Phenomena (e.g., photons, and the photoelectric effect) • Atomic Physics (e.g., the Rutherford and Bohr models, atomic energy levels, and atomic spectra) • Nuclear and Particle Physics (e.g., radioactivity, nuclear reactions, and fundamental particles) • Relativity (e.g., time dilation, length contraction, and mass-energy equivalence)	6-11	5-8
Miscellaneous		
• General (e.g., history of physics and questions of a general nature that overlap several major topics) • Analytical Skills (e.g., graphical analysis, measurement, and math skills) • Contemporary Physics (e.g., astrophysics, superconductivity, and chaos theory)	4-9	3-11
Skill Specifications: Physics		
Recall	Approx. % of Test	Approx. No. of Items
• Generally involves remembering and understanding concepts or information	20-33	15-25
Single-Concept Problem		
• Recall and use of single physical relationship	40-53	30-40
Multiple-Concept Problem		
• Recall and integration of two or more physical relationships	20-33	15-25

FIGURE 1. The percentage and number of test items by content and skill specifications for the SAT Subject Test in Physics. SAT materials selected from the SAT Subjects Tests Preparation Booklet, 2006. Reprinted by permission of the College Entrance Examination Board, the copyright owner.

remoteness from the detailed curricula taught in individual classrooms, feeds a public interest to focus on and monitor educational accountability. Large-scale tests, developed from models of test specifications, are the instruments of choice to measure critical but generalized knowledge and skills (Atkinson, 2004; College Entrance Examination Board, 2006).

As with the benefits, the costs associated with the model of test specifications are found in its simplicity and its broad sampling of behavior. Because the cognitive model of test specifications represents a "picking-and-choosing" of the most important content and skills in the program of study, diagnostic claims about examinees' strengths and weaknesses, either behavioral or cognitive, are not usually compelling. To defensibly support such claims would require intensive testing of particular content and skills in order to rule out alternate explanations for correct or incorrect answer choices. Moreover, the assessments derived from a model of test specifications do not usually include a critical mass of items designed to measure any particular skill. It is understandable that the intensive measurement of any one skill is usually absent from these tests because they are designed to measure many different behaviors within a short time. Of course, the consequence of this practice is that the test will not usually support diagnostic claims about students' strengths and weaknesses because content and skills are not tested with enough frequency or depth.

In addition to not testing any content and skill thoroughly, there is also an absence of empirical evidence to suggest that the content and skills outlined in the test specifications are being used by examinees. Aside from accepting the informed judgment of content experts, there are few empirical studies, if any, conducted to determine the nature of the thinking processes examinees actually use to solve and answer the test items of interest. We often accept this absence of empirical evidence in part because we find intuitive theories of mind to be quite plausible, particularly when developed by content experts. Perhaps we find it difficult to imagine what other content and skills (outside of the ones included in the test specifications) examinees might use to generate correct responses; as adults it is difficult to think as children or adolescents do. We find it difficult to imagine how examinees would answer items correctly by using an entirely different set of thinking processes from the one included in the test specifications. Certainly, the thinking processes used by examinees could not be entirely different from those outlined in the test specifications. But as long as the thinking processes examinees use

to answer test items are sufficiently different from the content and skills in the test specifications, the validity of diagnostic inferences will be open to question. When models of test specifications have been evaluated, an increasing number of studies have shown that the cognitive skills reflected in the test specifications failed to match important aspects of examinees' thinking (Gierl, 1997; Hamilton, Nussbaum, & Snow, 1997; Poggio et al., 2005).

The cognitive model of test specifications is a convenient, albeit simple, cognitive model of learning. However, this convenience comes at the cost of not being able to provide strong psychological evidence for diagnostic claims about examinees' thinking processes. Using a student's large-scale test performance to *predict* how that student will perform in college may be warranted by studies showing the predictive validity of the assessment. Yet, using an examinee's large-scale test performance to *explain* the thinking processes underlying that performance is unwarranted in the absence of empirical studies showing that examinees who answered specific classes of items did so correctly or incorrectly for the expected reasons. Although the information obtained from large-scale tests can be used for the purpose of rank-ordering examinees and providing information about some of the behavioral skills examinees have mastered, this information cannot be easily extended to make defensible diagnostic claims about their thinking processes.

The Cognitive Model of Domain Mastery

Description

The second cognitive model is of *domain mastery.* In its purest form, a model of domain mastery is generated to illustrate the population of knowledge and skills that is believed to conceptualize expertise or mastery within a circumscribed achievement domain. Curriculum-based tests or embedded assessments are representative of the cognitive model of domain mastery (Idol, 2007; Idol, Nevin, & Paolucci-Whitcomb, 1999; Shinn, 1998; Wilson & Sloane, 2000). Although there are different varieties of curriculum-based tests, these tests tend to have many similarities because they are often designed by teachers to thoroughly evaluate students' knowledge and skill achievements in a program of study at a specific grade level. Curriculum-based tests are used for the purpose of formative evaluation and can be described as an "academic thermometer" to assess student progress on a broad set of knowledge and skill components. As a formative evaluation tool, curriculum-based tests are standardized so that changes in test scores can be ascribed to changes in student progress rather than idiosyncrasies with the test itself.

Curriculum-based tests are designed to measure breadth and depth of knowledge and skills. In fact, the comprehensive focus of these assessments seems to have grown out of a general dissatisfaction with traditional large-scale tests (Idol et al., 1999; Wilson & Sloane, 2000), which are often summative and include only a sampling of the basic knowledge and skills in the domain of interest. Understandably, this sampling of basic knowledge and skills is perceived to create a potential mismatch between the standardized, large-scale test content and the full range of material actually taught by teachers and learned by students (Idol et al., 1999; Louis, Febey, & Schroeder, 2005; Wilson & Sloane, 2000). Although items in curriculum-based tests are also sampled from a universe of items, multiple tests are designed to be administered on multiple occasions so as to measure content and skills intensely, including all components of the curriculum that would otherwise be impossible to measure with a single test (Brown, Campione, Webber, & McGilly, 1992; Idol et al., 1999; Resnick & Resnick, 1992; Shinn, 1998; Wilson & Sloane, 2000). Although teachers and content experts might not view a comprehensive listing of knowledge and skills as representative of a cognitive model of learning, this listing does represent such a model. Test score inferences are validated based on the population of content (knowledge) and skills included in the curriculum-based test specifications. An examinee who performs successfully on a series of curriculum-based tests is assumed to have mastered the knowledge and skills within the domain of interest.

The sequence of steps normally involved in developing the test specifications for a series of curriculum-based tests illustrates the attention paid to a universe of knowledge and skills from which the tasks are drawn. For example, close attention is paid to the curricular cohesion or "conceptual flow" in the following 10 steps used for developing a series of mathematics tests at the elementary school level (Idol et al., 1999): (a) the table of contents of the course textbook is normally used to help define the scope of the knowledge and skills in the domain, including the sequence charts, placement tests, and review tests; (b) a list of relevant concepts is identified within the curriculum; (c) a raw-data sheet is constructed containing the list of relevant concepts along with their corresponding page numbers in the course textbook; (d) the concept list is reordered if the order is not progressive and logical; (e) all of the concepts are checked to ensure that they have had enough coverage; (f) complementary work materials are selected if the result of step (e) is found lacking; (g) the curriculum is reorganized by relevant concept, sequence, and textbook page number; (h) the concepts that could be and were taught simultaneously are identified and connected; (i) assessments, including placement tests and review tests, are developed to represent the organization of the conceptual flow identified; and (j) placement tests are administered to students. In condensed form, Figure 2 illustrates the resulting test specifications after invoking these 10 steps (the complete specifications

Sets and Numbers

	Subskill	Pages	Total Items
1.	sets	2,3	6
2.	counting	70, 71, 80, 81, …	83
3.	comparison	4, 5, 6, 7, …	172
⋮			
8.	odd-even	8, 9,…	27
9.	ordinal numbers	72, 73, 89T	24
10.	bar graphs	10, 11, 122, 147 17	

Place Value

	Subskill	Pages	Total Items
1.	hundreds, tens	66, 67, 68, 69,…	131
2.	thousands	194, 195, …	76
⋮			
4.	millions	202, 203	16
5.	expanded form	86, 89T, 199,…	44

Addition

	Subskill	Pages	Total Items
1.	basic facts 0-10	14, 15, 18, …	140
2.	basic facts 11-20	24, 25, 26,…	191
⋮			
8.	story problems	38, 39, 45T,…	52

FIGURE 2. The scope and sequence of concepts and skills considered for an elementary-level curriculum-based mathematics test. Adapted from *Models of curriculum-based assessment: A blueprint for learning*, 4th edition, by L. Idol, 2007. Austin, TX: PRO-ED, Inc. Copyright, 2007 by PRO-ED, Inc. Reprinted with permission.

would be many pages in length, and would include all knowledge and skills of interest). Multiple tests are constructed by generating items that reflect different topics and levels of the curriculum, and then administered on successive occasions.

Other examples of curriculum-based tests can be found in Shinn (1998), Wilson and Sloane (2000), and Wilson (2005). In particular, Wilson and Sloane (2000) present four principles guiding the Berkeley Evaluation and Assessment Research (BEAR) System. These principles include (a) adopting a developmental perspective, (b) matching instruction and assessment, (c) promoting teacher management and responsibility, and (d) gathering quality evidence. The principle of adopting a developmental perspective requires users of the Assessment System to shift their view of testing from a static activity to a dynamic activity, which "focuses on the process of learning and on an individual's progress through that process" (p. 183). The principle of matching instruction and assessment requires that the framework for the curriculum and tests be developed simultaneously so that each informs the other. Otherwise, tests may not represent the curriculum taught in the classroom. The third and fourth principles of the Assessment System are significant because they require teachers to cooperate with its new demands, and to oversee and maintain the technical quality of the testing instruments.

Evaluation

A benefit of the cognitive model of domain mastery is that it provides detailed information about the knowledge and skills required to achieve mastery of curricula. The value of the information obtained about examinee test performance is useful in so far as it provides a comprehensive account of examinees' *behavioral* outcomes. The tests are designed to probe multiple knowledge and skills thoroughly to ensure breadth and depth of mastery at a behavioral level. As a consequence, curriculum-based tests are designed to be sensitive to changes in student progress so that mastery can be confirmed or remediation can be implemented early to improve behavioral skills.

In spite of the strengths associated with the model of domain mastery, such as the emphasis on comprehensive learning outcomes, the curriculum taught in the classroom, and multiple opportunities for student feedback, there are limitations with the model. For example, assessment becomes time consuming. In order to generate detailed information about what students know, multiple tests need to be administered over multiple days to adequately assess all knowledge and skills thoroughly. Time is therefore an issue and may burden teachers and students. However, the main limitation with the cognitive model of domain mastery is its concentrated focus on *behavior*. The focus is largely on what behaviors a student should demonstrate, and less so, at a cognitive level, on how a student should process information in the service of the behavior. Mastery is defined largely according to behavioral outcomes (e.g., can the student add numbers?) and less so according to specific thinking processes (e.g., does the student use the correct strategy for adding the numbers?). Although multiple tests are designed to measure examinees' mastery of the knowledge and skills in the program of study, the thinking processes underlying this performance are often not explicitly measured and, therefore, not used to discriminate among distinct forms of mastery (see Barnett & Koslowski, 2002 for distinctions among experts based on thinking processes). Thus, curriculum-based test results may provide strong evidence for inferences about examinees' behavioral strengths and weaknesses but less convincing evidence for diagnostic inferences about their *cognitive* strengths and weaknesses.

Wilson and Sloane's (2000) description of the BEAR Assessment System provides a possible exception, however, showing that in principle thinking processes can be measured. In the BEAR System, a majority of the assessment tasks are designed to be open-ended to permit examinees to explain their reasoning, and allow teachers to peer into the thinking structure underlying examinees' responses. This is an excellent feature of the assessments. Wilson and Sloane (2000) presented instances of open-ended questions implemented in the IEY (Issues, Evidence, and You) science course. These open-ended questions permitted examinees to show their scientific reasoning in support of their knowledge of scientific concepts, such as designing and conducting investigations, evidence and trade offs, communicating scientific information, and group interaction. The open-ended responses were then graded using the IEY scoring guides, which incorporate a common logic adapted from the SOLO taxonomy by Biggs and Collis (1982). These scoring guides involve exemplars of examinee answers at different levels of competency (e.g., complete and correct response, partially complete response, response with one correct aspect, and incorrect response with no correct aspects). The IEY scoring guides were used to rate examinees' open-ended responses according to their overall reasoning (i.e., objective reasons for response and evidence) and correctness of response. However, the scoring guides were not used to rate features of reasoning, such as depth of understanding or severity of misconceptions. To obtain such detailed diagnostic information, the scoring guides would probably have to include distinctions in epistemological thinking processes (Kuhn, 2001). One can imagine, for example, two students receiving identical ratings for a complete and correct response and yet having different reasoning paths of varying sophistication.

Upon closer inspection, then, there is a danger of treating students' reasoning as a basic behavioral outcome within the cognitive model of domain mastery. In the absence of a tangible framework for interpreting reasoning paths or thinking processes in terms of what such processes reveal about the level of mastery and learning, diagnostic inferences about students' cognitive strengths and weaknesses will be limited. Of course, this limitation arises not with the BEAR System itself but with the cognitive model underlying it—namely, the model's lack of an empirically-based psychological framework for identifying, categorizing, and interpreting examinee responses (however, see Briggs, Alonzo, Schwab, & Wilson, 2006, for an empirical framework). Without such a framework, it is unclear how a teacher should use the test results to help the student improve his or her thinking in relation to the tasks within a domain. How would a teacher know to recognize a misconception and not simply treat it as a case of incomplete reasoning? Without this framework, examinee test performance does not seem to inform instruction in a new way. As a result, even if a curriculum-based test includes the

measurement of reasoning processes, the test may not live up to its full potential in diagnosing learning if the model underlying it fails to include a defensible framework for evaluating aspects of examinees' thinking processes.

The Cognitive Model of Task Performance

Description

The third cognitive model is of *task performance*. In its purest form, a model of task performance is generated to illustrate the thinking processes underlying the knowledge and skills students apply *in vivo* when solving educational tasks in a specific domain. One approach to generating a model of task performance is to administer a task or set of tasks to a group of students that is representative of the population of interest, and employ standard think-aloud methods with the intent to conduct protocol or verbal analysis (Chi, 1997; Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, in press-b; Newell & Simon, 1972; Taylor & Dionne, 2000). This model is the type that educational or cognitive researchers develop to empirically confirm the thinking processes individuals use to answer or solve classes of test items.

Currently, tests based on cognitive models of task performance are relatively uncommon in educational measurement. The absence of these types of tests in comparison to large-scale and curriculum-based tests indicates the relative newness of developing measures based on empirically-based cognitive models (NRC, 2001). However, there are some notable exceptions; namely, some of the diagnostic types of assessments generated from Embretson's (1998, 2002, 2005) Cognitive Design System, Mislevy's Evidence Centered Design (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004; Mislevy et al., 2003), and Tatsuoka's Rule Space Model (Tatsuoka, Corter, & Tatsuoka, 2004). To illustrate a cognitive model of task performance, we will consider Embretson's (2002, 2005) development of abstract reasoning items using an empirically-based theory of matrix processing that received solid evidential support from studies of protocol analysis, eye movement, computer simulations, and analyses of error rates (Carpenter, Just, & Shell, 1990).

Illustrated in Figure 3, a matrix completion problem is designed to measure a person's reasoning skill within abstract and unfamiliar domains (Raven, Court, & Raven, 1992). The thinking processes underlying this essential skill have been hypothesized to help individuals induce patterns in learning domains (Garlick, 2002). Carpenter et al. (1990) developed a theory of how individuals of varying ability manipulate and use information while solving tasks similar to matrix completion problems. Carpenter et al.'s theory includes two general thinking processes required to reason about and solve these problems. First, the process of finding a correspondence between figures in the problem (noticing abstract relationships among relevant attributes); and second, the process of inducing a rule to describe the correspondences or relationships between figures (identifying a type of rule that describes all the identified associations). The latter thinking process makes a heavy demand on the central executive in working memory, which monitors goal attainment in problem-solving tasks (see Carpenter et al., 1990 for a full description of the theory).

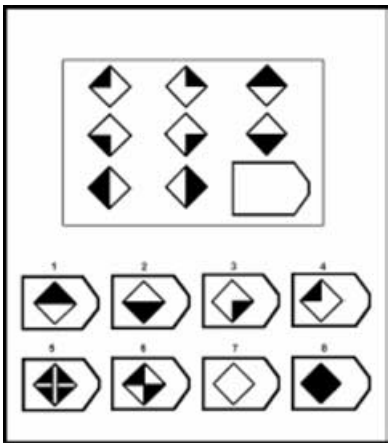

FIGURE 3. An example of a matrix completion problem used to measure fluid intelligence originally illustrated in Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431. Reprinted with permission.

According to Carpenter et al., individuals solve matrix completions problems by engaging these two general processes (finding correspondences and inducing rules) and a series of more specific thinking processes illustrated by the model of task performance shown in Figure 4. In this model, individuals solve matrix completion problems by following a sequence of steps: (a) encoding the figures of the first two entries in a row; (b) determining corresponding elements; (c) comparing the attributes of corresponding elements; (d) inferring a rule instance that applies to describing the relationship between corresponding elements; (e) encoding the third entry; and (f) comparing its corresponding elements with the other entries in the row and again inferring a rule. Problem solvers repeat these six steps for the first row of a matrix problem, followed by the second row, and then the rule that is inferred for rows 1 and 2 is mapped to the third row. One of the main difficulties for individuals attempting to solve matrix completion problems is finding a rule that describes the correspondences among figural attributes. Carpenter et al.'s theory suggests that individuals consider their rule options (i.e., constant in a row, quantitative pair wise progression, figure addition or subtraction, distribution of three values, distribution of two values) serially, selecting the rule that best describes the association among attributes only after other rules have been reviewed and discarded.

In an effort to generate abstract reasoning test items based on an empirically-based cognitive model, Embretson (2002)[3] used Carpenter et al.'s theory of matrix processing. Embretson first used hierarchical regression analysis to mathematically model the difficulty of matrix completion problems according to the characteristics outlined in Carpenter et al.'s theory. In particular, problems were identified as varying in difficulty along the following variables: abstract correspondence, the number of rules required for solution, and their joint effect on memory load. In support of Carpenter et al.'s theory, Embretson found that these item variables were indeed strong predictors of item difficulty (with $R = .74$, $p < .01$ for multiple models). Embretson then developed abstract reasoning items feature by feature using Carpenter et al.'s theory. To this end, Embretson demonstrated how a formal item structure, in which "slots" holding variations of identified item features (e.g., degree of abstract correspondence, number of rules), could be developed to successfully generate abstract reasoning items of varying cognitive complexity and difficulty.

Carpenter et al.'s theory in the design of abstract reasoning items is reflective

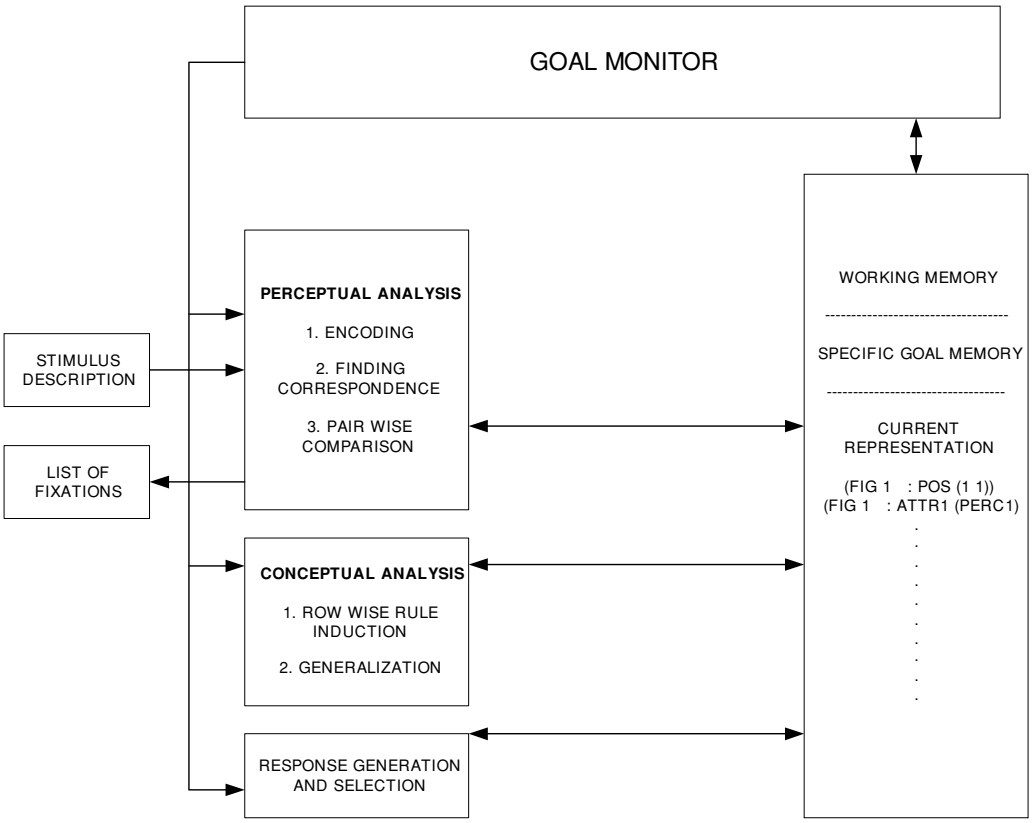

FIGURE 4. A cognitive model of processing of matrix completion items originally illustrated in Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431. Reprinted with permission.

of a cognitive model of task performance for educational measurement purposes. The model is based on a theory that has been empirically tested with a sample of the population of interest (college students), predicts performance for a task of interest (abstract reasoning items), and explains problem solving on the task according to specific thinking processes. The model is therefore useful in supporting diagnostic inferences about examinees' abstract reasoning processes. Although some might argue that abstract reasoning test items do not represent standard educational items but, rather, psychological tasks in the service of measuring fluid intelligence, this should not undermine the illustrative goal of the example. Not only is performance on standardized educational test items and psychological tasks correlated (e.g., Raven's Advanced Progressive Matrices), but both tasks are believed to invoke common cognitive processes for their solutions (Embretson & Gorin, 2001; Lohman, 2000; Stanovich, Sá, & West, 2004; Sternberg, 1999).

Evaluation

The benefits of developing test items from a cognitive model of task performance rest with the specificity and defensibility of the diagnostic claims that are subsequently generated about examinees' thinking processes. This specificity is possible because test items are developed based on detailed accounts of how students manipulate information and think through problems within a domain of interest. As such, the cognitive model of task performance must be sufficiently detailed in its preliminary organization to inform predictions about students' thinking and problem solving on a set of items (see Ericsson & Simon, 1993) where falsification of the model is possible with different populations or within different circumstances. In its final organization, the model must be adequately detailed to inform test item development so as to measure the component processes examinees are expected to use to solve educational items. The defensibility of the claims made from tests generated from a model of task performance is achieved on two fronts: first, when the model informing the test is generated and validated with studies involving students, claims about examinee test performance are grounded in empirical evidence; second, when the model informing the assessment is tested against alternate models of task performance, in which one model is shown to systematically predict student performance, claims about examinees' thinking processes are unlikely to be attributable to other models.

The specificity and defensibility of the information generated from tests based on a model of task performance come at a cost, however. The model of task performance will often illustrate a narrow set of knowledge and skills because the thinking processes underlying each set must be sufficiently detailed to offer description and prediction of student performance, as well as guide item development. If tests informed by models of task performance are used to make specific, diagnostic claims about examinees' problem solving, then test items should be developed

to measure isolated component processes, whenever possible (although see Behrens et al., 2004, for an example where component processes may be extracted from constructed-response tasks). Items measuring relevant component processes simultaneously may confound diagnostic information associated with individual processes, and fail to provide unequivocal evidence for cognitive strengths and weaknesses.

There are few, if any, achievement tests based on cognitive models of task performance. This absence may occur for a number of reasons. First, the narrowness of the set of knowledge and skills included in a model of task performance may make assessments based on this model incompatible with the targets of inference usually desired. Second, the cognitive model of task performance requires psychological evidence from the populations to which inferences will be made. Obtaining this kind of empirical evidence is expensive, requiring extensive research time and human resources. Third, it may be the case that test developers' expertise is still considered to be an adequate substitute, at far less expense, for formal models of task performance. History is an important consideration since many well-known achievement tests were developed before the interest and perceived necessity of blending educational measurement and cognitive psychology. Yet, once a model of task performance is generated and tested against alternate models, test developers can exercise maximum control in measuring specific thinking processes in examinees.

Implications for Practice

In this section we discuss the blending of the three cognitive models for the development of test items that are maximally informative in supporting diagnostic claims about examinees' thinking processes. As mentioned at the beginning of the paper, our interest in describing the different cognitive models in their purest form was not to suggest that the models could not or should not be blended to improve educational tests. Instead, our interest was to first identify and clarify the current state of models commonly used in educational measurement. It is our supposition, based upon strong assertions from well-known and respected agencies such as the U.S. Department of Education and the OECD, that there is ambiguity as to the kinds of cognitive models (and tests) that do indeed support claims about examinees' cognitive strengths and weaknesses. We suspect that these assertions are not anomalies but very likely reflect systemic ideas that other researchers and practitioners may also hold about the diagnostic information that can be reasonably extracted from educational tests.

In the foregoing sections of the paper, we have argued that the cognitive models of test specifications and domain-mastery are not often grounded in empirically-based investigations of student thinking and performance. A cognitive model of test specifications is, in its purest form, a model that represents a set of informal beliefs about examinees' knowledge and skills. The model of test specifications is often used to generate items for large-scale tests with the purpose of measuring a broad sample of knowledge and skills. These knowledge and skills, however, may not represent the expected thinking processes examinees actually use to respond to test items. Diagnostic inferences gleaned about examinees' cognitive strengths and weaknesses from large-scale tests will therefore be limited.

The model of domain mastery in its purest form illustrates the comprehensive knowledge and skills associated with behavioral competence within an academic domain. Although the behavioral competence measured is assumed to also reflect cognitive competence, the empirical studies show this link are often missing. The model of domain mastery can be a highly informative *behavioral* model of what knowledge and skills students should display in the domain of interest. Curriculum-based tests developed from this model will support specific claims about examinees' behavioral strengths and weaknesses. Teachers and/or curriculum experts can develop this category of model without needing to conduct empirical studies in its support. However, since the model is not supported with psychological studies of how students generate knowledge, reason, or know how to apply specific skills, diagnostic inferences about examinees' cognitive strengths and weaknesses will be limited.

Although we have identified the cognitive model of task performance as the sole model that can defensibly support diagnostic inferences about students' thinking processes, it is important to recognize that we have not suggested that this is an option without serious limitations and trade-offs. For example, the model on its own is narrow in the knowledge and skills it represents. This characteristic occurs because the thinking processes underlying the knowledge and skills represented in the model must be illustrated explicitly. This level of detail is a necessity since the model must inform the development of test items that can support specific diagnostic claims about examinees' thinking processes, including the mechanisms and pathways that delineate performance. Also, model development will be time consuming because it must be generated with empirical evidence from human participants and, ideally, also tested against alternate models.

In an ideal world, all three categories of models (and tests) would be "blended" to reap the benefits of each. One example of achieving the best of many worlds is to have a type of curriculum-based, large-scale test that focused on measuring examinees' thinking processes. This type of test could be informed by *many* cognitive models of task performance, with each model of task performance focused on illustrating the component processes underlying a narrow range of knowledge and skills (see also ECD framework, Mislevy et al., 2003). The collective set of models of task performance would be nested within a broader model of test specifications or model of domain mastery. When taken together, these models would represent the thinking processes underlying a *wide* range of knowledge and skills. This type of test, created from an array of cognitive models, would not necessarily have to be summative. Instead, multiple tests, each measuring increasingly more sophisticated knowledge and skills, could be administered throughout the school year with the objective of informing student learning and instruction.

When models of test specifications, domain mastery, and task performance are considered collectively in the design of an assessment, their function is similar to the models described in Robert Mislevy's Evidence-Centered Design or ECD (Mislevy et al., 2003). For example, within ECD, the *assembly model* could be compared to the model of test specifications; the *student*

Table 1. Dimensions Used to Evaluate the Value of Information Provided by Three Cognitive Models for Educational Measurement

	Information			
Cognitive Model/ Assessment	**Content Coverage/ Range of Skills**	**Depth of Knowledge and Skills Measured**	**Psychological Evidence**	**Focus or Assessment Goals**
Test Specifications/ Large-Scale Assessments	Moderate	Low	Low	Rank ordering or behavioral mastery
Domain Mastery/ Curriculum-Based	High	High (behavioral)	Low	Behavioral strengths and weaknesses
Task Performance/ Cognitive Diagnostic	Low	High (cognitive)	High	Cognitive strengths and weaknesses

model could be compared to the model of domain mastery; and the *task model* could be compared to the model of task performance. Finally, the *evidence model* in ECD could be viewed as an executive model designed to oversee the coordination of the other three models. The comparisons of the models described in the present paper with the models outlined in ECD should be viewed as a way to draw parallels with other systems of thought and is not meant to suggest that they are identical. In particular, the model of task performance described in the present paper requires a substantial commitment to collecting empirical evidence of the response processes examinees use to answer classes of test items.

Returning from an ideal world to our own imperfect one, in which most tests are developed from models of test specifications or domain mastery, only observations about examinees' behavioral responses (whether they answered an item correctly or incorrectly) are viable. Even if these behavioral responses are invoked by underlying thinking processes, diagnostic inferences about examinees' cognitive strengths and weaknesses based on models of test specifications and domain mastery are not defensible. If the purpose of the test is to generate diagnostic inferences about examinees' cognitive strengths and weaknesses, an empirically substantiated cognitive model of task performance should be used to develop the test. This is the case not because the model of task performance is inherently superior to the models of test specifications and domain mastery, but because this is the only model that requires empirical evidence of the thinking processes underlying the knowledge and skills measured by the test.

There are also obstacles with blending the three categories of models in practice. Although it is beyond the scope of this paper to enumerate all these obstacles, we do identify some of the more obvious ones. For example, large-scale tests could, in principle, be developed from a combination of cognitive models involving test specifications, domain mastery, and task performance. However, when one considers the breadth of many large-scale tests on the one hand and the depth of the model of task performance on the other, it is not immediately clear that the large-scale tests and models of task performance should or could easily partner up. This partnership can, of course, be forced by modifying the time permitted for developing and administering large-scale tests—or, alternatively, by changing the empirical rigor of the model of task performance. But is this worth it? In trying to satisfy too many testing purposes, could we not be essentially undermining them all with testing tools that try to do too much but, in the end, do nothing very well?

Governments and testing companies do have the resources to invest in the development of large-scale tests that can deliver diagnostic information about cognitive strengths and weaknesses. To be sure, a development program to create such large-scale tests will require extensive time and human resources up front so as to determine the nature of the knowledge and skills to be modeled and measured. Several different types of cognitive models have been developed and applied to large-scale reading comprehension items in order to mine these items for what they reveal about examinees' strengths and weaknesses (Buck, Tatsuoka, & Kostin, 1997; Embretson & Wetzel, 1987; Gorin, 2005). That these models are distinct in their characteristics, some involving cognitive processes and others involving exclusively behavioral characteristics, is valuable for identifying the best models for accounting for student performance. However, only tests designed from models of thinking processes can support diagnostic inferences about examinees' *cognitive* strengths and weaknesses.

Conclusion

We have illustrated three categories of cognitive models in educational measurement. The critical characteristics of these models are illustrated in Table 1 and include the range of knowledge and skills represented, the depth of the knowledge and skills measured, and the extent of psychological evidence for the inferences generated. These characteristics help define the boundaries for the kinds of diagnostic claims that can be defensibly made about examinees based on their test performance. For example, if the target of inference is to rank examinees within a domain or judge their basic competence, then a model of test specifications is satisfactory. If the target of inference is to judge examinees' behavioral strengths and weaknesses within a domain, then a model of domain mastery is satisfactory. However, if the target of inference is to evaluate examinees' cognitive strengths and weaknesses, then a model of task performance is best used to develop the test. Although we have identified three categories of cognitive models, we are confident other categories may exist to cater to specific assessment needs.

The integrated use of these cognitive models in educational measurement has the potential to give educational tests a genuine and useful role for

diagnosing students' strengths and weaknesses (Leighton & Gierl, in press-a). As tools in identifying content understanding and misunderstanding, educational tests developed from the collective consideration of these models, especially the model of task performance, stand to provide specific cognitive information about where in the learning process students succeed and fail. However, we do not want to conclude the paper with a vague assertion of how the consideration of cognitive models is desirable. Before we can generate diagnostic claims about students' thinking processes, it is necessary to recognize the dominant models informing test development and what these models can and cannot do for supporting defensible claims about student cognition. The upshot of what we have attempted to articulate is that unless there is genuine commitment to fully developing all three categories of models—especially the model of task performance—inferences about examinees' cognitive strengths and weaknesses are only as good as the primary cognitive model underwriting such claims. Such news may be discouraging but we think realistic. We hope it provides an incentive for developing and evaluating more cognitive models of task performance, as well as the psychometric procedures to support them.

Notes

[1]We are evaluating cognitive models only in terms of their capacity to inform specific and defensible claims about students' cognitive strengths and weaknesses. To this end, the models are examined for their capacity to (a) inform test item development for measuring cognitive processes of interest, and (b) support the construct validity of inferences made about students' cognitive processing.

[2]These cognitive models could be compared to the models described in the Evidence-Centered Design (ECD) system (Mislevy, Steinberg, & Almond, 2003).

[3]Embretson (2002, 2005) normally refers to these cognitive psychological models as *processing* or *componential processing models.* Processing or componential processing models (also *generative models*, Bejar, 2002) serve the same function as the model of task performance in so far as they are generated and validated using empirical evidence of students' cognitive processing on the test of items of interest.

References

Anderson, J. R., Qin, Y., Sohn, M. H., Stenger, V. A., & Carter, C. S. (2003). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin & Review*, *10*, 241–261.

Atkinson, R. C. (2004). Achievement versus aptitude in college admissions. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions*. New York: Routledge.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge, UK: Cambridge University Press.

Barnett, S. M., & Koslowski, B. (2002). Adaptive expertise: Effects of type of experience and the level of theoretical understanding it generates. *Thinking and Reasoning*, *8*, 237–267.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global E-learning program. *International Journal of Testing*, *4*, 295–301.

Bejar, I. (2002). Generative testing: From conception to implementation. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives, Handbook 1: Cognitive domain*. Addison Boston: Wesley.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice item. *Educational Assessment*, *11*, 33–63.

Brown, A. L., Campione, J. C., Webber, L. S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments* (pp. 121–212). Boston: Kluwer Academic.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*, 423–466.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*, 271–315.

College Entrance Examination Board (2006). *SAT Subject Tests™ Preparation Booklet*.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 629–660). Chichester, England: John Wiley.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (2005). Measuring human intelligence with artificial intelligence. In R. J. Sternberg and J. E. Pretz (Eds.), *Cognition and Intelligence* (pp. 251–267). Cambridge, UK: Cambridge University Press.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*, 175–193.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Garlick, D. (2002). Understanding the nature of the general factor of intelligence: The role of individual differences in neural plasticity as an explanatory mechanism. *Psychological Review*, *109*, 116–136.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research*, *91*, 26–32.

Gierl, M. J., Tan, A., & Wang, C. (April, 2005). *Identifying content and cognitive dimensions on the SAT* (Research Report No. 2005-11). New York: The College Board.

Gorin, J. S. (2005) Manipulation of processing difficulty on reading comprehension test questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351–373.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *Spring*, 17–27.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, *10*, 181–200.

Healy, A. F. (Ed.) (2005). *Experimental cognitive psychology and its applications*. Washington, DC: American Psychological Association.

Idol, L. (2007). *Models of curriculum-based assessment: A blueprint for learning* (4th

ed.). Austin, TX: Pro-Ed International Publisher.

Idol, L., Nevin, A., & Paolucci-Whitcomb, P. (1999). *Models of curriculum-based assessment: A blueprint for learning* (3rd ed.). Austin, TX: Pro-Ed International Publisher.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Kalchman, M., Moss, J., & Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. In S. M. Carver and D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 1–38). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Kuhn, D. (2001). How do people know? *Psychological Science*, *12*, 1–8.

Lane, S. (2004). 2004 NCME Presidential Address. Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, *23*, 6–14.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *Winter*, 1–10.

Leighton, J. P., & Gierl, M. J. (Eds.) (in press-a). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (in press-b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton and M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236.

Leighton, J. P., & Gokiert, R. (2005). *The cognitive effects of test item features: Identifying construct irrelevant variance and informing item generation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). Cambridge, UK: Cambridge University Press.

Louis, K. S., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis*, *27*, 177–204.

Luecht, R. M. (2005). *Extracting multidimensional information from multiple-choice question distractors for diagnostic scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

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

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

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education*, *2*, 283–308.

Organization for Economic Cooperation and Development (OECD) (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

Pearson, P. D., & Garavaglia, D. R. (2003). *NAEP Validity studies: Improving the information value of performance items in large scale assessments*, NCES 2003–08. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, *24*, 307–353.

Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. (2005). *Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills*? Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's progressive matrices and vocabulary scale*. San Antonio, TX: Psychological Corporation.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–76). Boston, MA: Kluwer Academic.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*, 234–247.

Rogers, W. T., & Yang, P. (1996). Testwiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*, 247–259.

Shinn, M. R. (Ed.) (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford Press.

Siegler, R. S. (2005). Models of categorization: What are the limits? In L. Gershkoff-Stowe and D. Rakison (Eds.), *Building object categories in developmental time* (pp. 433–439). Mahwah, NJ: Lawrence Erlbaum Associates.

Snow, R. (1993). Construct validity and constructed-response tests. In R. Bennett and W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Stanovich, K. E., Sá, W. C., & West, R. F. (2004). Individual differences in reasoning. In J. P. Leighton and R. J. Sternberg (Eds.), *Nature of reasoning* (pp. 375–409). New York: Cambridge University Press.

Sternberg, R. J. (1984). What cognitive psychology can and cannot do for test development. In B. S. Plake and J. Mitchell (Eds.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 39–60). Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology*, *24*, 359–375.

Tatsuoka, K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*, 901–926.

Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *92*, 413–425.

Thissen, D., & Edwards, M. C. (2005). *Diagnostic scores augmented using multidimensional item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

United States Department of Education (2004). Testing for results: Helping families, schools, and communities understand and improve student achievement. In *NCLB* (*Stronger accountability*) (chap. 1). Retrieved February 15, 2006, from http://www.ed.gov/nclb/accountability/ayp/testingforresults.html.

Wang, X. B., Deng, H., Williams, K., & Laitusis,

V. (2005). *Assessing the interpretability of four diagnostic reporting methods for the new SAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.
Wellman, H. M., & Lagattuta, K. H. (2004). Theory of mind for learning and teaching: The nature and role of explanation. *Cognitive Development*, *19*, 479–497.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*, 181–208.