Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: National Council on Measurement in Education and American Council on Education.

Leighton, J. P., & Gierl, M. J. (in press). *Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes*. Educational Measurement: Issues and Practice.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694 (Monograph Suppl. 9).

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1–103). New York: American Council on Education/Macmillan.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erlbaum.

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

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Upper Saddle River, NJ: Prentice Hall.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575–603.

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

No Child Left Behind Act of 2001, Pub.L. No. 107–110, 115 Stat. 1435 (2002).

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: Author.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243–299). Hinsdale, IL: Dryden Press.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187–214.

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.

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

U.S. Department of Education. (2004, September 16). *Stronger accountability: Testing for results: Helping families, schools, and communities understand and improve student achievement*. Retrieved February 15, 2006, from http://www.ed.gov/nclb/accountability/ayp/testingforresults.html

2

The Demand for Cognitive Diagnostic Assessment

Kristen Huff and Dean P. Goodman

In this chapter, we explore the nature of the demand for cognitive diagnostic assessment (CDA) in K–12 education and suggest that the demand originates from two sources: assessment developers[1] who are arguing for radical shifts in the way assessments are designed, and the intended users of large-scale assessments who want more instructionally relevant results from these assessments. We first highlight various themes from the literature on CDA that illustrate the demand for CDA among assessment developers. We then outline current demands for diagnostic information from educators in the United States by reviewing results from a recent national survey we conducted on this topic. Finally, we discuss some ways that assessment developers have responded to these demands and outline some issues that, based on the demands discussed here, warrant further attention.

THE DEMAND FOR COGNITIVE DIAGNOSTIC ASSESSMENT FROM ASSESSMENT DEVELOPERS

To provide the context for assessment developers' call for a revision of contemporary assessment practices that, on the whole, do not operate within a cognitive framework, we offer a perspective on existing CDA literature, and we outline the differences between psychometric and cognitive approaches to assessment design. The phrases *working within*

[1] The term *assessment developers* is used here to refer to psychometricians, cognitive psychologists, curriculum and instruction specialists, and learning scientists who are practitioners and/or members of the assessment research community.

a cognitive framework, cognitively principled assessment design, and *cognitive diagnostic assessment* are used interchangeably throughout this chapter. They can be generally defined as the joint practice of using cognitive models of learning as the basis for principled assessment design and reporting assessment results with direct regard to informing learning and instruction.

Perspective and Overview of Cognitive Diagnostic Assessment Literature

One way to portray the history of CDA is to start with Embretson's (1983) publication in *Psychological Bulletin*, where she effectively integrated advances in cognitively psychology and contemporary notions of construct validation: "construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge stores that are involved in performance" (p. 180). Although cognitive psychologists had been working for quite some time modeling the relationship between item difficulty and cognitive processes (Fischer & Formann, 1982), Embretson's publication was significant in its application of developments from cognitive psychology to measurement theory. Messick (1989) notes Embretson's contribution in his hallmark chapter on test validity:

> As cognitive psychology in general and information-processing models of cognition in particular have advanced over the years, producing powerful experimental and quantitative techniques of task decomposition, this modeling approach has become much more salient in measurement circles. Although in one form it has only recently been incorporated into the validity literature under the rubric of construct representation (Embretson, 1983), the explicit probing of the processes productive of performance has long been part of the validation repertoire. (Cronbach, 1971; Cronbach & Meehl, 1955, p. 27)

In the same volume of *Educational Measurement*, others challenged the educational measurement community to reconceptualize testing theory and practice to reflect student cognition more broadly than a single latent trait (Snow & Lohman, 1989), and to better integrate testing with instruction and learning (Nitko, 1989). These chapters voiced a clear demand for more discourse among obvious collaborators: educational measurement specialists (psychometricians) and cognitive psychologists. Since the late 1980s, several researchers, theorists, and practitioners echoed the demand for more cognitively informed test design, scoring, and reporting to better inform teaching and learning (e.g., Bennett, 1999; Chipman, Nichols, & Brennan, 1995; Feltovich, Spiro, & Coulson, 1993; Mislevy, 1996; National Research Council [NRC], 2001; Pellegrino, Baxter, & Glaser, 1999). In response, many promising complex, cognitively based scoring models have been proposed, such as Tatsuoka's (1983, 1995) rule-space method; DiBello, Stout, and Roussos' (1995; see also Roussos et al., this volume) unified model; and Leighton, Gierl, and Hunka's (2004; Gierl, Leighton, & Hunka, this volume) attribute hierarchy method. In addition, much research has been devoted to developing innovative (e.g., performance-based, computer-based) item types that purportedly measure higher-order thinking skills (e.g., Bennett, 1999; Bennett & Bejar, 1998; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997; Sireci & Zenisky, 2006). Item difficulty modeling research has built on Embretson's work in task decomposition and construct representation (Embretson, 1999; Gorin, 2005; Sheehan, 1997; Sheehan, Ginther, & Schedl, 1999). Last, one of the most persistent themes to emerge from this area is how assessments developed from a psychometric perspective differ substantially from assessments designed within a cognitive framework. Much of the *demand* for CDA originates from discussion about the potential to inform teaching and learning by changing the way in which we design assessments.

Psychometric versus Cognitive Assessment Design Frameworks

Nichols (1993, 1994) paints a picture of two starkly different approaches to assessment design and use of assessment results in his discussion of tests developed from psychometric versus cognitive frameworks. Psychometrically developed tests are primarily developed by evaluating the statistical properties of items because the chief purpose of these assessments is to rank-order examinees along a highly reliable scale (which implies the measurement of a largely unidimensional construct) for the purpose(s) of selection, classification, and/or summative evaluation. Such an approach is contrasted with assessments developed from within a cognitive framework, which are primarily used to diagnose the learning state of the examinee and to inform remediation. Nichols (1994) and others (NRC, 2001; Pellegrino et al., 1999; Snow & Lohman, 1989) have compellingly argued that educational assessments designed from psychometric models are not optimal for informing instruction because the assessment tasks were not designed from an explicit model of how students learn, and scoring models that are primarily used to rank-order examinees are severely limited in their ability to reflect the complexity of

the learner's cognitive strengths and weaknesses. Consequently, the test results are not necessarily connected to classroom learning and instruction, and accordingly, have limited utility for educators and students.

Researchers have argued that to maximize the educational benefits from assessments, these exams should be situated within an aligned and integrated system of curriculum, instruction, and assessment (Nichols, 1993, 1994; NRC, 2001; Pellegrino et al., 1999). In this system, *curriculum* should sequence learning objectives that reflect our understanding of how students build knowledge structures and expertise in the specified domain, *instruction* should employ strategies that facilitate knowledge building and active learning, and *assessment design* should be informed by the same cognitive framework that shapes the curriculum and should provide feedback to teachers that informs instruction. That is, learning and instruction are optimized when a cognitive model of learning not only provides the framework for assessment design, but also provides the framework for the educational system in which the assessment is used.

Pellegrino (2002) elaborates this integrated system of curriculum, instruction, and assessment by suggesting a cognitive assessment framework that consists of three interrelated elements:

- A model of student learning in the specified academic domain
- A set of beliefs (or hypotheses) about the kinds of observations that will provide evidence of student competencies in the domain, where such competencies are defined by the cognitive model
- A framework for interpreting the results of the assessment

This general cognitive assessment framework is operationalized through evidence-centered design (Steinberg et al., 2003) or, as it has been more recently described, principled assessment design (Mislevy & Riconscente, 2005). Principled assessment design is an approach to designing assessment tasks that are explicitly linked theoretically and empirically to the targets of measurement through the use of detailed design templates (Riconscente, Mislevy, & Hamel, 2005). Design templates require specific attention to what kind of knowledge is being measured, how the target of measurement is related to proficiency in the domain, how the various assessment task features are related to different levels of proficiency, and how the scoring model supports valid interpretations about student proficiency. Such precise articulation and

documentation of the assessment argument facilitates a transparency in test design that is too often missing from exams developed from a psychometric perspective.

Some may argue that assessment design practices, whether working from within a cognitive framework or a psychometric framework, are essentially the same because psychometrically developed tests require that items are written to test specifications, and such specifications typically span both content and "cognitive" skills, such as problem solving, reasoning, analysis, and application. However, this position is not necessarily correct. In practice, test specifications for large-scale assessments often only specify content requirements and give little or no explicit consideration to the types of cognitive skills that underlie a curriculum. For example, in our review of test development material from several state departments of education Web sites (Massachusetts Department of Education, 2004; New Jersey Department of Education, 2006; Washington State Office of Superintendent of Public Instruction, 2006), we found general references to the assessment of various cognitive skills, but no explicit application of these skills in each state's test specifications. The lack of any explicit consideration to cognitive skills in the development of items and test forms for a state assessment has also been reported in the work of O'Neil, Sireci, and Huff (2004).

In contrast to psychometrically developed tests that do not explicitly assess cognitive skills, principled assessment design ensures that the cognitive skills of interest are explicitly targeted during item and test form development. This explicit targeting of cognitive skills has three key benefits. First, it helps ensure that all relevant cognitive skills are considered during item and test form development, and that the test forms assess an appropriate balance of cognitive skills (something that cannot be ensured by psychometrically developed tests that only explicitly consider content). Second, it helps ensure that the rationales supporting task designs are clearly documented (something that would facilitate transparency and would give assessment developers some useful evidence to support the validity of their assessments). Third, and perhaps most important, it helps ensure that the resulting scores lead to meaningful and valid interpretations about students' cognitive skills and abilities.

The research supporting the instructional potential of principled assessment design within a cognitive framework is convincing and growing. Since 1999, the NRC has commissioned three compelling volumes

that summarize the vanguard in this area. In 1999, *How People Learn: Brain, Mind, Experience, and School* outlined the latest developments in the science of learning and how learning environments can be designed to take advantage of what we know about how students build proficiency. *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001) followed soon after and addressed how educational assessment can be designed to reflect modern principles of learning by embodying five principles: designing assessments from a model of cognition and learning, making explicit the relationships among task design elements and the cognitive model, collecting evidence for the cognitive processes elicited by the tasks (i.e., construct validation), considering score reporting issues at the beginning of the assessment design phase, and ensuring comparability and fairness across all learner groups (p. 176). Then, in 2002, *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools* was published. In this volume, the authors argued that the cognitive-based instructional and assessment principles outlined in the two previously mentioned volumes were the most promising guidelines for improving science and mathematical achievements in secondary education.

Although assessment developers' demands for the implementation of CDA principles are compelling, do their demands reflect the needs of the primary users of assessment data – classroom teachers? To what extent are teachers making use of diagnostic information that is currently available from large-scale assessments, and how could this information be improved to better meet teachers' needs? We explore these questions next.

THE DEMAND FOR COGNITIVE DIAGNOSTIC ASSESSMENT FROM EDUCATORS

When assessing the demand for CDA from educators, it is important to recognize that they are *not* actually demanding that assessment developers use cognitive models as the basis for assessment design and reporting. What educators are demanding is that they receive instructionally relevant results from any assessments in which their students are required to participate and that these assessments be sufficiently aligned with classroom practice to be of maximum instructional value. In this section, we explore this demand by highlighting results from a national survey conducted by Goodman and Huff (2006) that examined U.S.

teachers' beliefs about, and practices relating to, the use of diagnostic information from state-mandated and commercial large-scale assessments.

For the purposes of the survey, *state-mandated* assessments were defined as standardized tests that states require schools to administer at specific grade levels (e.g., tests that satisfy the terms of the No Child Left Behind [NCLB] Act of 2001). In most cases, these state-mandated assessments would be developed by or for the state, but they could also include commercial assessments that are administered to satisfy state and federal accountability requirements. Typically, these assessments are designed to assess student proficiency in relation to state curriculum standards, and the resulting scores are reported as proficiency levels (e.g., Basic, Proficient, or Advanced). Proficiency level results are typically defined by general descriptions of the types of knowledge and skills students in each level possess (e.g., students at the Proficient level "demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems" [Massachusetts Department of Education, 2005, p. 9]) or by more detailed, subject-specific descriptions (e.g., eighth-grade mathematics students at the Proficient level are able to "predict from data displays; apply measures of central tendency; describe patterns and relationships using algebraic equations" [Missouri Department of Elementary and Secondary Education, 2005, p. 15]).

Commercial large-scale assessments were defined as standardized tests, such as the Iowa Test of Basic Skills (ITBS), California Achievement Test, and the Stanford Achievement Test, that schools and districts may administer to groups of students for their own *local* use (i.e., assessments not required by the state), or tests such as the Preliminary Scholastic Aptitude Test (PSAT)/National Merit Scholarship Qualifying Test (NMSQT), Scholastic Aptitude Test (SAT), or American College Test (ACT) that students may take in preparation for college admission. Typically, these commercially available tests are norm-referenced, and the feedback consists primarily of various subscores and nationally based percentile ranks.

The survey respondents were a nationally representative random sample of 400 elementary and secondary mathematics and English language arts teachers in U.S. public and nonpublic schools. Results of this survey are estimated to be within 4.90 percentage points of the true population value 95% of the time.

How Many Teachers Receive Large-Scale Assessment Results?

With the implementation of NCLB (2001) and the proliferation of testing and results-based accountability in both public and nonpublic schools, one should expect that the number of teachers in the United States who receive large-scale assessment results is high. The findings of our survey bear this out, showing that the vast majority of elementary and secondary mathematics and English language arts teachers receive results from state-mandated and commercial large-scale assessments. As shown in Figure 2.1, of the 400 teachers who participated in the survey, only 5% reported that they did not receive assessment results from state-mandated assessments, and only 7% reported that they did not receive results from commercial large-scale assessments. Only 3% of teachers reported that they did not receive results from either one of these types of large-scale assessments.

How Often Do Results from Large-Scale Assessments Inform Instruction?

Although it appears that most mathematics and English language arts teachers in the United States have access to large-scale assessment results, to what extent are these results being used? Data we have collected suggest that large-scale assessment results are being used, although not always as regularly as assessment developers and policy makers would like and, perhaps, believe.

In our survey, 45% of teachers who received state-mandated assessment results indicated that these results inform their instruction daily or a few times a week, and 81% stated that these results informed their instruction a few times a year or more (see Figure 2.2 for a more detailed breakdown of these results). A much lower, but still substantial, percentage (27%) of teachers who received commercial large-scale assessment results indicated that these results inform their instruction daily or a few times a week, and 68% of teachers who received these results stated that these results inform their instruction a few times a year or more (see Figure 2.2 for a more detailed breakdown of these results).

Given the stakes that are attached to many large-scale assessment results today, the use of these results by large proportions of teachers is not surprising. What is surprising, however, is that significant proportions of teachers who receive large-scale assessment results appear to *never* use them to inform their instruction or do so only *once* a year. As shown in Figure 2.2, 14% of teachers who received results from

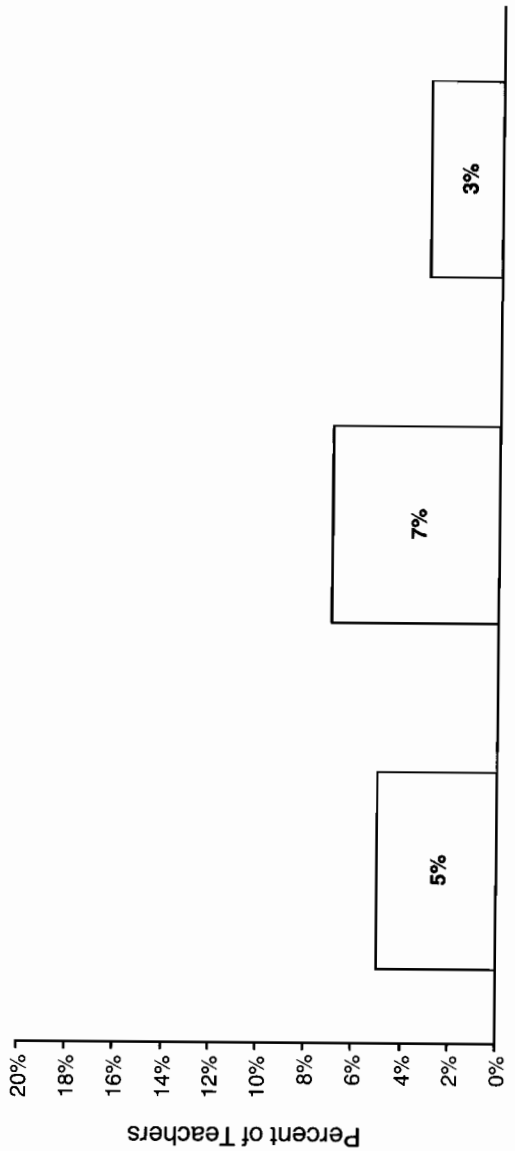

FIGURE 2.1. Percentage of elementary and secondary mathematics and English language arts teachers who do not receive state-mandated and/or commercial large-scale assessment results.

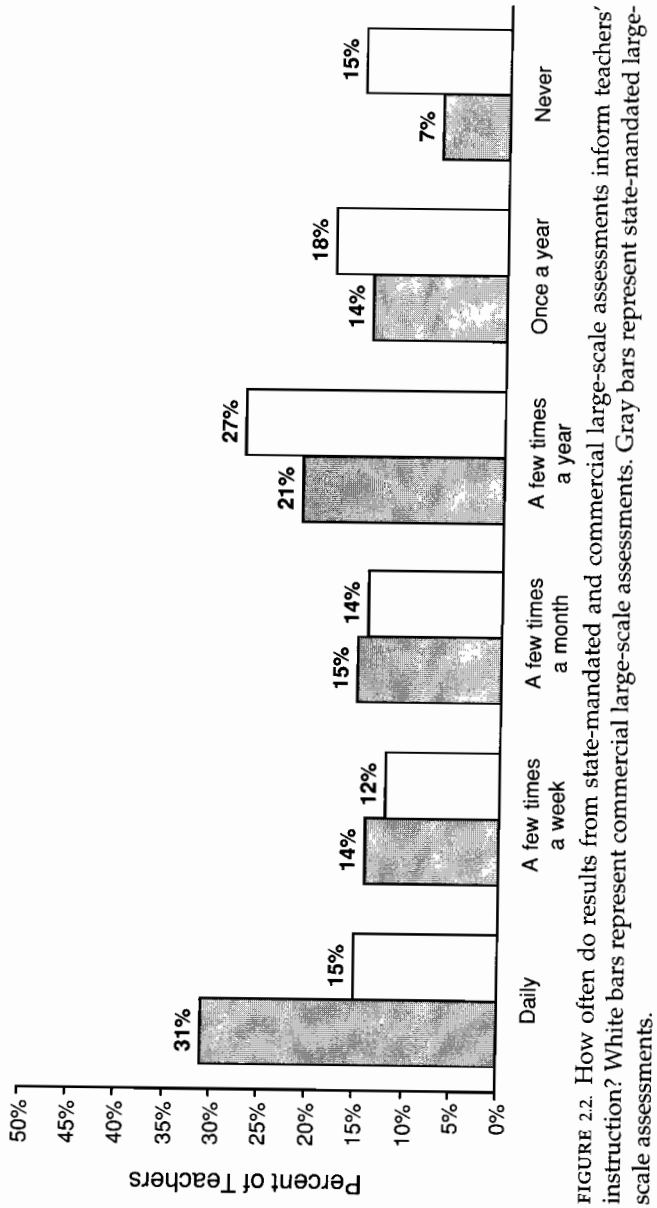

FIGURE 2.2. How often do results from state-mandated and commercial large-scale assessments inform teachers' instruction? White bars represent commercial large-scale assessments. Gray bars represent state-mandated large-scale assessments.

state-mandated assessments reported that they use these results only once a year, and 7% never use these results at all. The findings for commercial assessments should be an even greater cause of concern for assessment developers: 18% of teachers who received results from commercial assessments reported that they use these results to inform their instruction only once a year, and 15% reported that they do not use these results at all (see Figure 2.2).

Although it is encouraging to see that many teachers are using large-scale assessment results to inform their instruction on a regular basis, our findings suggest that far too many teachers who currently receive these results never use them, or use them only once a year for their ultimate purpose – to inform instruction. Clearly, a more concerted effort by assessment developers is needed to make these results more instructionally relevant and useful to all teachers, especially to those teachers who receive the results but rarely or never use them, presumably because they do not find the results to be relevant to their instruction.

To What Extent Do Teachers Believe It Is Appropriate to Collect Diagnostic Information Using a Variety of Assessment Methods?

Teachers have an assortment of assessment options that can inform their instructional practice. As part of our research, we were interested in finding out to what extent teachers believe it is appropriate to collect diagnostic information using a variety of assessment methods.

Based on our experience working with K–12 teachers, we were not surprised to find that the vast majority (93%) of teachers believed that it is moderately appropriate or very important to collect diagnostic information using assessments produced by the classroom teacher (see Figure 2.3F). Clearly, most teachers believe that the assessments they develop are well suited for identifying students' specific strengths and weaknesses. Indeed, the results in Figure 2.3 suggest that teachers believe classroom assessments are the *best* way to collect diagnostic information (with 59% of teachers stating that assessments produced by the classroom teacher are *very* appropriate for collecting diagnostic information).

Given their potential for being ongoing, integrated with instruction, and tailored to the specific needs of individual students, it is hard to deny that well-designed classroom assessment practices (especially those that are designed for formative purposes) can offer valuable diagnostic information. Unfortunately, however, not all assessment practices used by classroom teachers are well designed. Indeed, considerable research

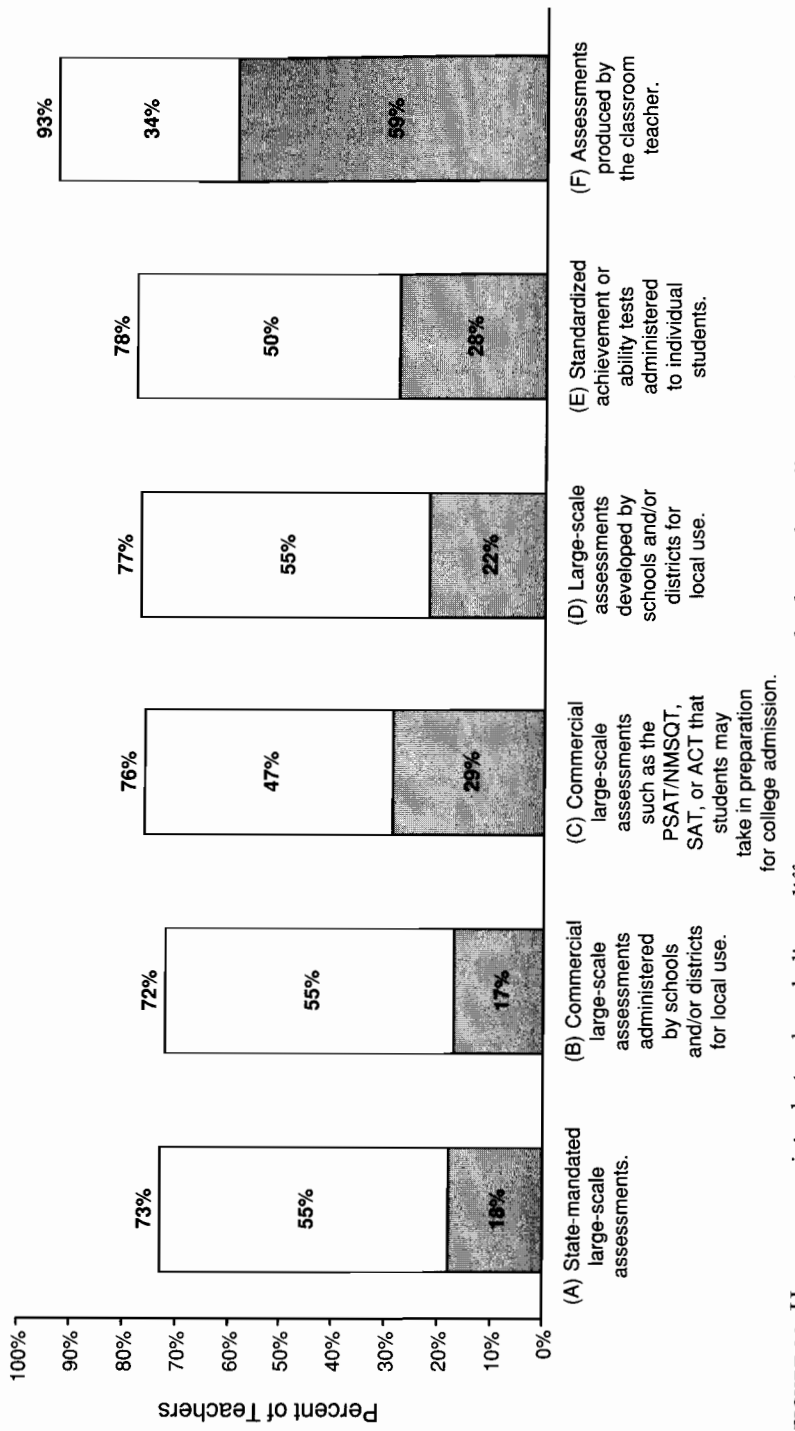

FIGURE 2.3. How appropriate do teachers believe different assessment methods are for collecting diagnostic information? White bars represent the percentage of teachers who think the methods are moderately appropriate for collecting diagnostic information. Gray bars represent the percentage of teachers who think the methods are very appropriate for collecting diagnostic information.

exists on the various ways in which classroom assessment practices can be improved (see Black & Wiliam, 1998a, 1998b; Stiggins, 2001), including better specifications of the types of cognitive skills and abilities that should and will be assessed (Mathematical Sciences Education Board, 1993; Notar, Zuelke, Wilson, & Yunker, 2004). Efforts to help teachers incorporate CDA principles into the design of classroom assessment practices would therefore appear to have some clear benefits.

In addition to classroom assessments, high percentages of teachers believed it is also appropriate to collect diagnostic information using state-mandated and commercial large-scale assessments. Seventy-three percent and 72% of teachers, respectively, believed it is moderately appropriate or very appropriate to collect diagnostic information using state-mandated large-scale assessments or commercial large-scale assessments such as the ITBS or Stanford Achievement Test (see Figures 2.3A and 2.3B). Seventy-seven percent of teachers believed that it is moderately appropriate or very appropriate to collect diagnostic information using large-scale assessments developed by schools and school districts for local use (see Figure 2.3D). Surprisingly, 76% of teachers had a similar view about collecting diagnostic information from college admission assessments such as the SAT or ACT (see Figure 2.3C). These results are especially noteworthy given that only a slightly higher percentage of teachers (78%) shared similar views about the appropriateness of collecting diagnostic information using standardized achievement or ability tests administered to individual students (the latter of which are typically designed and used for the purpose of providing diagnostic information; see Figure 2.3E). Based on these findings, it appears that any efforts to collect diagnostic information using large-scale assessments would be consistent with the general views of most teachers.

What Types of Information Do Teachers Consider Diagnostic?

Under NCLB (2001, Section 111[b][3][c][xiii]), states are required to produce diagnostic score reports that allow parents, teachers, and principals to understand and address the specific needs of students. We believe that this requirement is a step in the right direction because it emphasizes the need to make score reports both meaningful to the intended audience and directly relevant to addressing student needs. Unfortunately, the legislation does not provide any guidance as to what kinds of information are considered diagnostic and how this requirement could be best accomplished.

We found that very high percentages of teachers consider existing types of large-scale assessment results to be diagnostic (see Figure 2.4). Seventy-two percent of teachers surveyed considered even the most basic type of large-scale assessment results, overall subject-level scores, to be diagnostic information. Eighty-seven percent of teachers considered subdomain scores to be diagnostic information. Eighty-five percent and 82% of teachers, respectively, considered descriptions of specific skills or knowledge a student demonstrates on a large-scale assessment and descriptions of specific skills or knowledge that a student should develop to be diagnostic information. Eighty percent of teachers considered item-level results to be diagnostic information. These results indicate that teachers are quite willing to regard large-scale assessment results as being diagnostic; the extent to which they believe that these results actually serve to inform instructional practice is explored next.

Does Diagnostic Information from Large-Scale Assessments Play a Valuable Role in Informing Instruction?

In our survey, we explored the roles that diagnostic information[2] from state-mandated and commercial large-scale assessments play in informing instruction. In almost all cases, a majority of teachers indicated that diagnostic information from state-mandated and commercial large-scale assessments plays a valuable role in informing instructional practices at all levels of the K–12 school system: individual student, classroom, grade, school, and district (see Figures 2.5 and 2.6). The only exception to this pattern was for commercial assessments, where less than half (46%) of the teachers reported that diagnostic information from commercial large-scale assessments plays a valuable role in informing instruction at the individual student level.

The results in Figures 2.5 and 2.6 also show that more teachers believe that diagnostic information from state-mandated assessments plays a valuable role in informing instruction at each level of the K–12 school system than diagnostic information from commercial large-scale assessments. The percentage of teachers who agreed or strongly agreed with statements that diagnostic information from state-mandated assessments plays a valuable role in instruction at different levels of the K–12 school system ranged from 59% at the individual student level to 79% percent at the classroom level (see Figure 2.5). The percentage

[2] *Diagnostic information* was operationally defined for these survey items as any information provided by a large-scale assessment that is more detailed than overall subject-level scores.

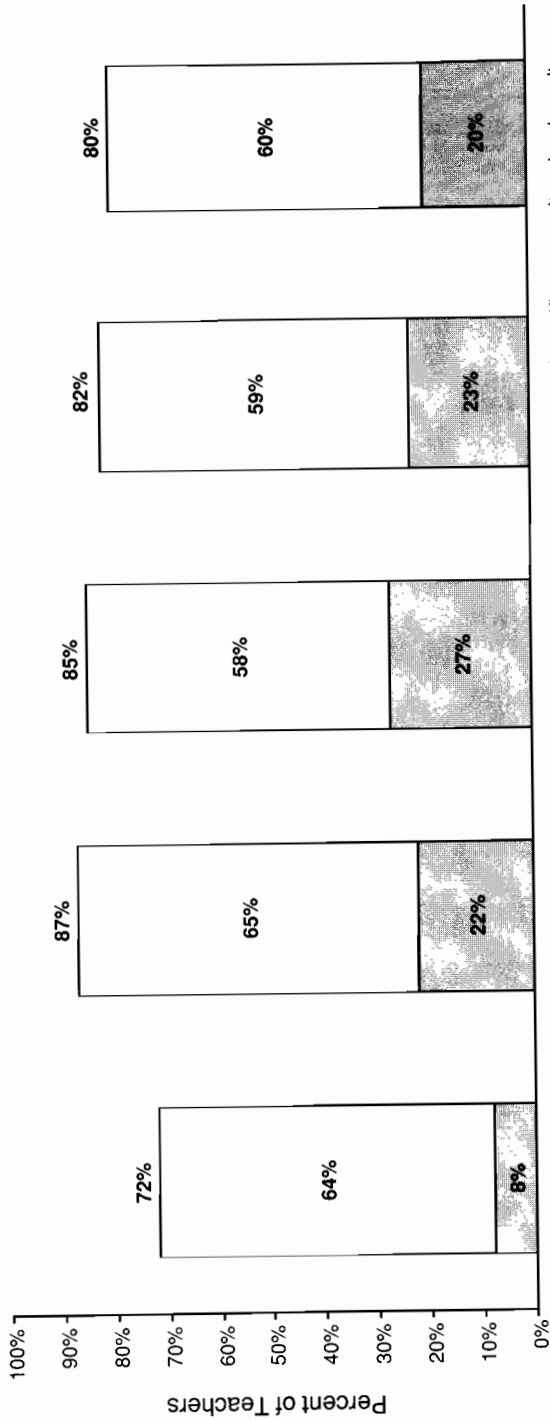

FIGURE 2.4. Extent to which teachers believe that different types of large-scale assessment results are diagnostic information. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

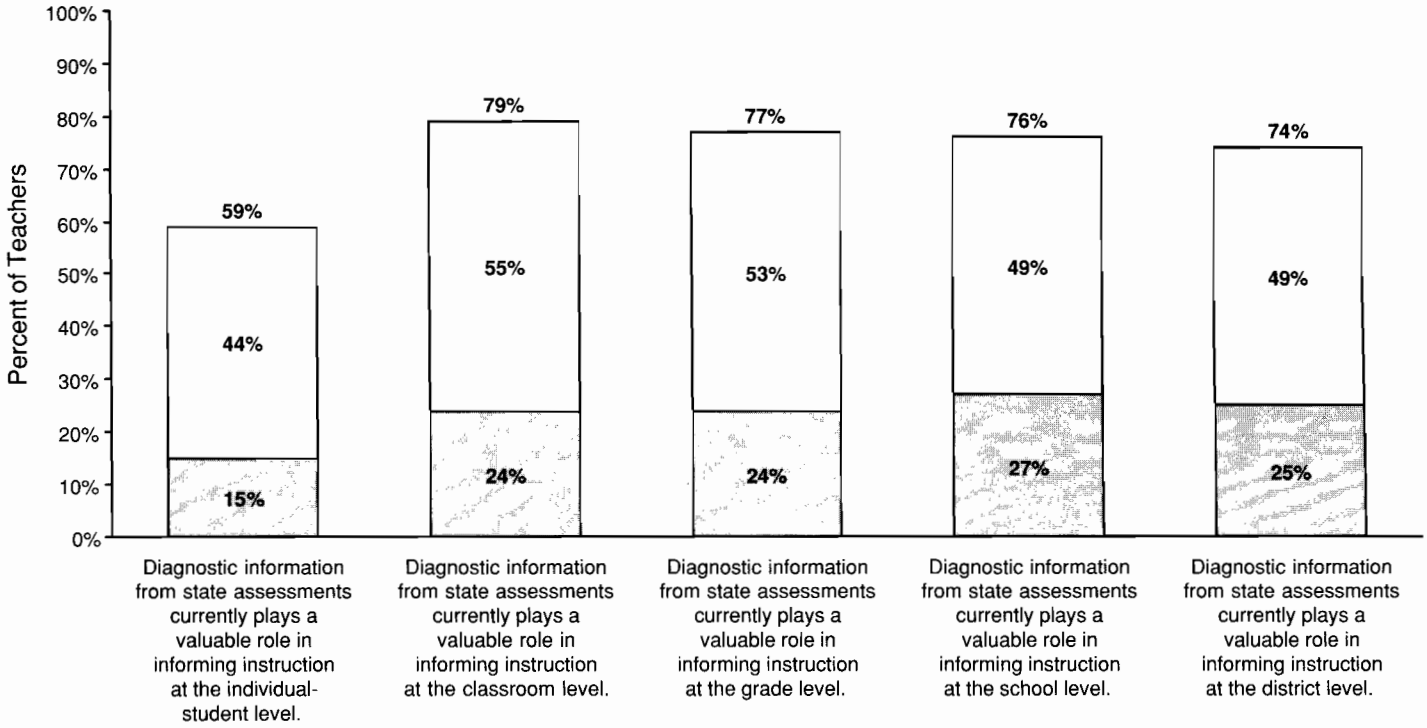

FIGURE 2.5. Extent to which teachers believe that diagnostic information from *state-mandated* large-scale assessments currently plays a valuable role in informing instruction at the individual student, classroom, grade, school, and district levels. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

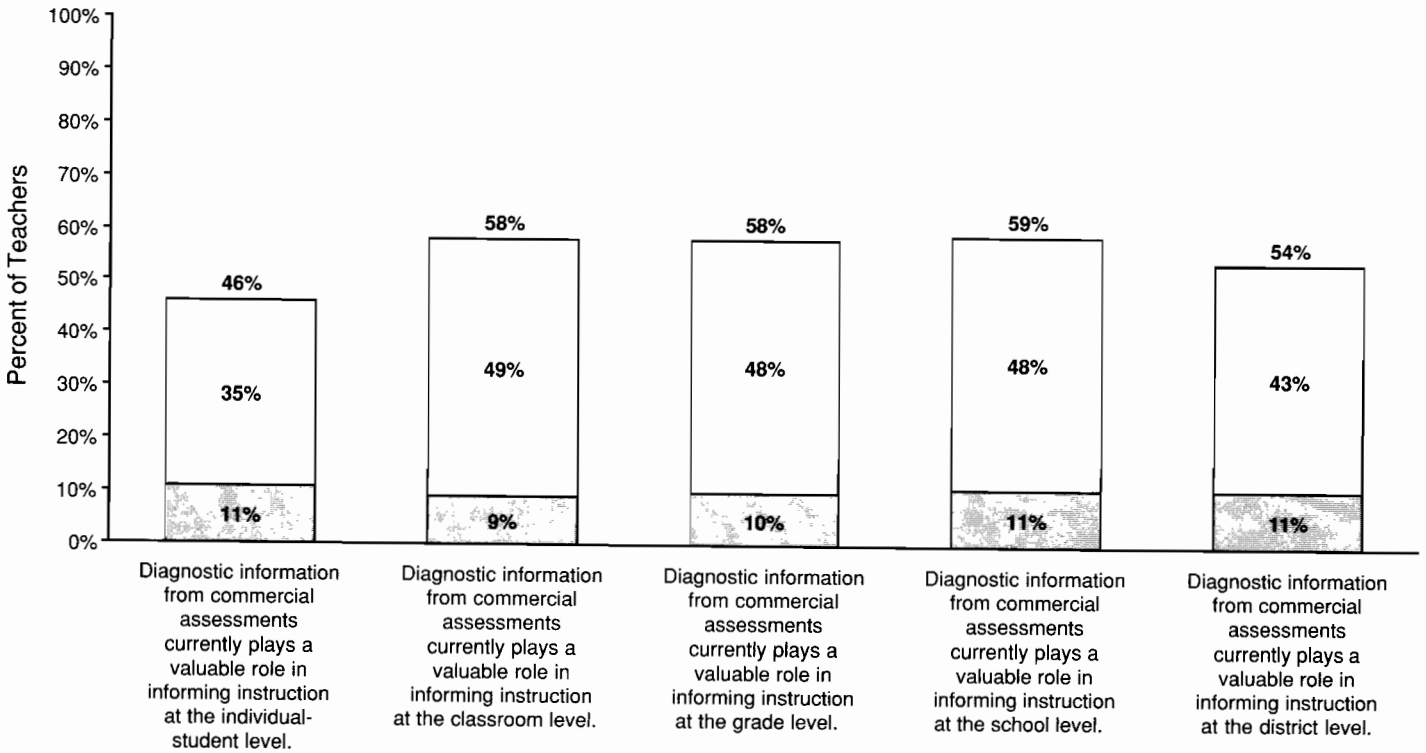

FIGURE 2.6. Extent to which teachers believe that diagnostic information from *commercial* large-scale assessments currently plays a valuable role in informing instruction at the individual student, classroom, grade, school, and district levels. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

of teachers who shared a similar view toward diagnostic information from commercial large-scale assessments was 13% to 21% lower across the various levels, ranging from 46% at the individual student level to 59% at the school level (see Figure 2.6). Based on these results, it appears that teachers believe that diagnostic information from state-mandated and commercial large-scale assessments have *less* instructional value at the individual student level than at the classroom, grade, school, and district levels. One can infer from these results that classroom teachers are more likely to use assessment results to inform general instructional practices rather than informing instruction with individual students.

How Is Diagnostic Information from Large-Scale Assessments Being Used by Teachers?

To help improve the utility of diagnostic assessment results, it is important to determine how the results are being used by teachers who receive them. As part of our research, we explored how often teachers use diagnostic information from state-mandated and commercial assessments across a variety of educational activities. Some of our findings are presented in Table 2.1.

Teachers reported similar patterns in their use of diagnostic information from state-mandated assessments and from commercial assessments, although information from commercial assessments is used less frequently overall. Approximately one-third (36%) of teachers who receive results from state-mandated assessments indicated that they use diagnostic information from these assessments regularly (i.e., daily or a few times a week) when planning their instruction and selecting instructional strategies. The percentage of teachers who use diagnostic information from commercial large-scale assessments to plan instruction and to select instructional strategies dropped by half to 18% and 17%, respectively. Although these findings suggest that some teachers make regular use of diagnostic information from large-scale assessments, given the intent of diagnostic feedback and the objectives of NCLB (2001), it is important to note that significant percentages of teachers who receive state-mandated assessment results *never* use them to inform instructional planning (13%), select instructional strategies (18%), assess their teaching effectiveness (20%), give feedback to students (22%), evaluate student progress (17%), or remediate students (29%). The percentages almost double when the same questions are asked about the

TABLE 2.1. *How often teachers use diagnostic information from state-mandated and commercial large-scale assessments across a variety of educational activities*

	% of teachers who use results daily or a few times a week		% of teachers who never use results	
	State-mandated assessments	Commercial assessments	State-mandated assessments	Commercial assessments
Planning my instruction	36	18	13	32
Selecting instructional strategies	36	17	18	33
Assessing my teaching effectiveness	21	10	20	39
Giving feedback to students	18	9	22	38
Remediating students	18	8	29	49
Evaluating student progress	16	7	17	38

use of commercial assessment results, where between 32% and 49% of teachers report that they *never* use diagnostic results to inform these activities. Overall, these results suggest that the teachers' current use of diagnostic information is limited, and that further effort is required to make assessment results from both state-mandated and commercial assessments more suitable for informing these types of important educational activities.

Obstacles That Inhibit the Use of Diagnostic Information from Large-Scale Assessments

Although it is encouraging that a majority of teachers believe that diagnostic information plays a valuable role in informing instructional practices, it is clear that substantial percentages of teachers never use diagnostic information for key educational activities. To help address the issue of nonuse of diagnostic information, it is important to consider what obstacles inhibit teachers' use of this information. Findings relevant to this issue are presented in Figures 2.7 and 2.8 and are discussed in this section.

Teachers reported that the most significant obstacles in using diagnostic assessment information were (a) not receiving the assessment results back in time to use them and (b) a lack of resources to inform the proper use of diagnostic information from large-scale assessments. In our survey, 68% of teachers reported not getting results back in time as an obstacle that inhibits the use of diagnostic information from state assessments, and 57% of teachers reported the same obstacle for using diagnostic results from commercial assessments (see Figures 2.7A and 2.8A). Approximately half (50% and 49%) of the teachers considered a lack of resources to inform the proper use of results as an obstacle that inhibited the use of diagnostic information from state-mandated assessments and commercial assessments, respectively (see Figures 2.7G and 2.8G). Given these high percentages, efforts by assessment developers to reduce the amount of time required to release assessment results and to provide classroom teachers with more resources to inform the proper use of diagnostic information from large-scale assessments are clearly warranted.

Approximately one-fourth of teachers (27% for state-mandated assessments and 24% for commercial assessments) indicated that the diagnostic information that is currently reported for large-scale assessments is not useful (see Figures 2.7D and 2.8D). Approximately one-third

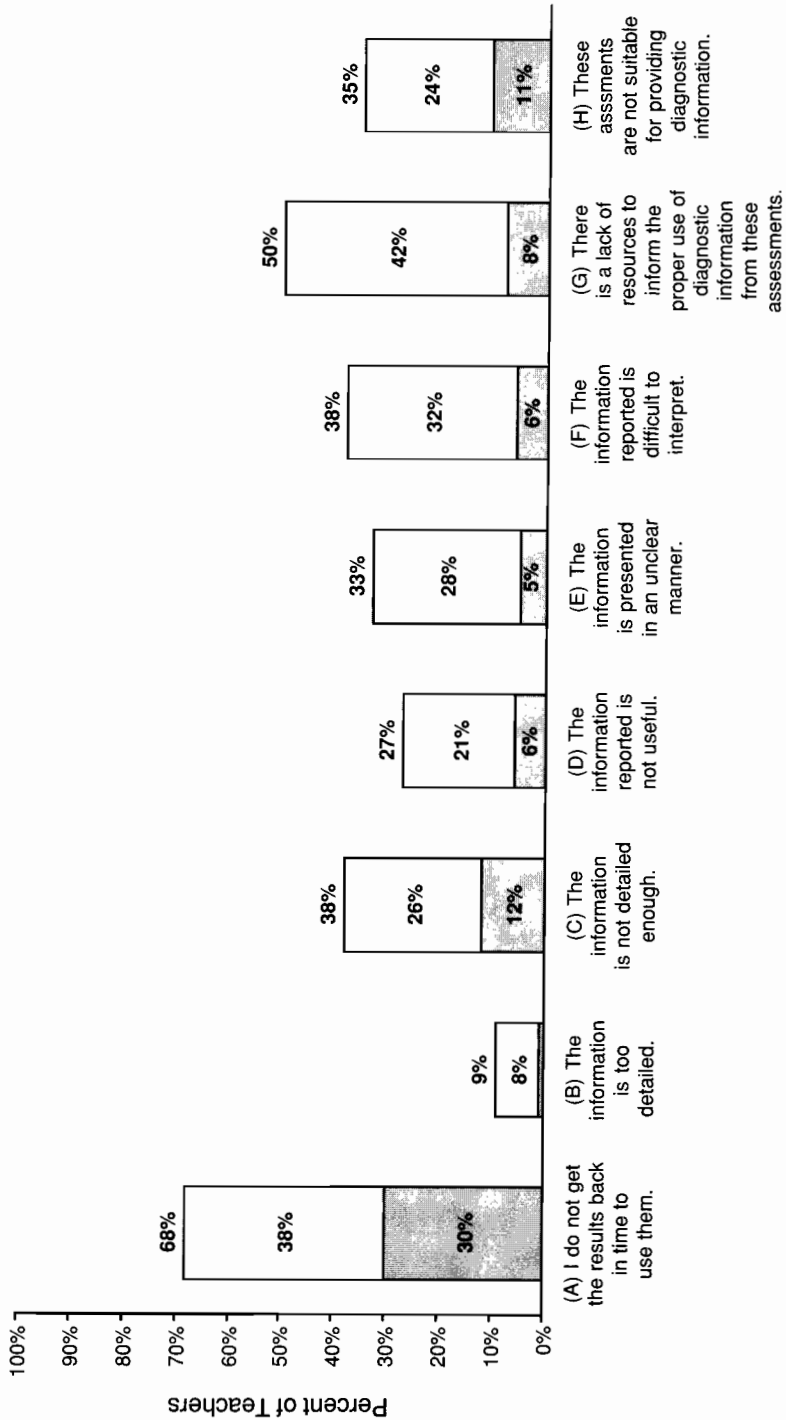

FIGURE 2.7. Obstacles that inhibit the use of diagnostic information from *state-mandated* large-scale assessments. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

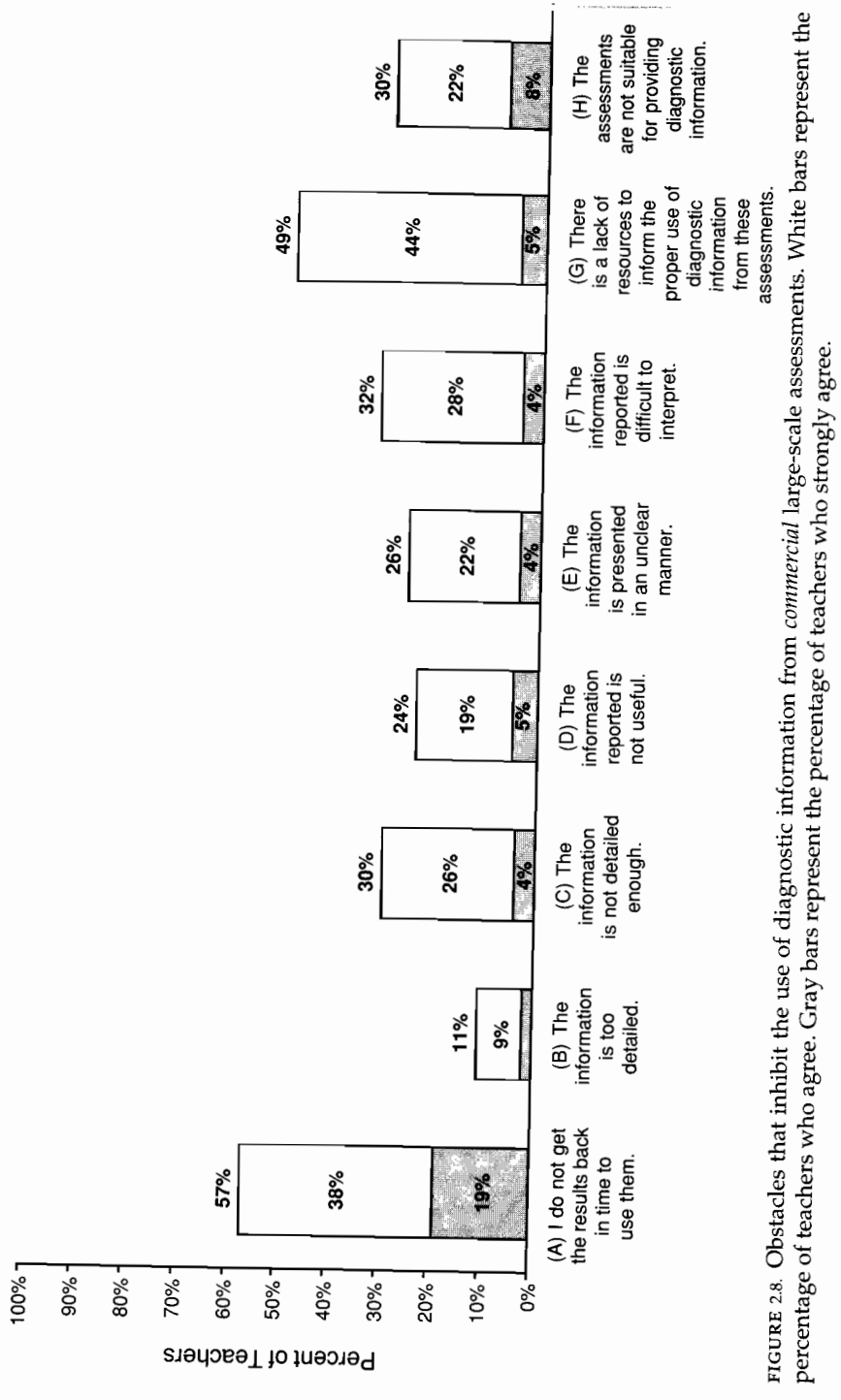

FIGURE 2.8. Obstacles that inhibit the use of diagnostic information from *commercial* large-scale assessments. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

of teachers (38% for state-mandated assessments and 32% for commercial assessments) also reported that the information reported on state-mandated and commercial large-scale assessments is difficult to interpret (see Figures 2.7F and 2.8F). Thirty-three percent of teachers believed that diagnostic information from state-mandated assessments is presented in an unclear manner (see Figure 2.7E); 26% of teachers had similar views regarding diagnostic information from commercial large-scale assessments (see Figure 2.8E). Based on these results, further efforts to make diagnostic information from large-scale assessments more useful to teachers and easier for all teachers to understand appear to be necessary.

In a more positive light, assessment developers should be encouraged to see that a majority of teachers do not believe that large-scale assessments are unsuitable instruments for providing diagnostic information. Sixty-five percent of teachers disagreed or strongly disagreed with the statement that "[State-mandated assessments are] not suitable for providing diagnostic information" (results derived from Figure 2.7H). Seventy percent of teachers who receive commercial assessment results disagreed or strongly disagreed with a comparable statement about commercial large-scale assessments (results derived from Figure 2.8H). These results suggest that most teachers regard large-scale assessments as an appropriate vehicle for collecting diagnostic data and are in keeping with results discussed previously. Still, considerable percentages of teachers (35% and 30%) agreed or strongly agreed with this statement for state-mandated and commercial assessments, respectively, so further efforts to better align large-scale assessments with their intended use appear to be warranted.

Do Large-Scale Assessment Results Currently Providing Sufficient Information about Students' Strengths and Weaknesses?

Another important finding of our survey was that a majority of classroom teachers believe that large-scale assessment results do not provide sufficient amounts of information regarding students' strengths and weaknesses. As shown in Figure 2.9, 51% of teachers believed that state-mandated large-scale assessments results *do not* provide sufficient information about students' strengths and weaknesses, and 53% of teachers held similar beliefs about commercial large-scale assessment results. Also, as shown in Figure 2.10, 74% and 71% of teachers believed that it would be valuable to have more diagnostic information than is typically

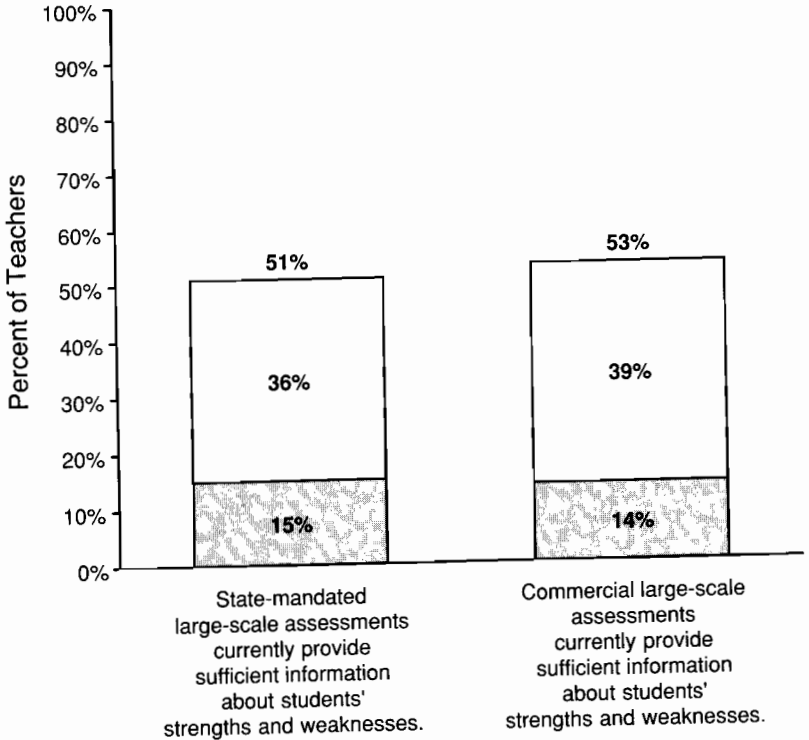

FIGURE 2.9. Extent to which teachers *do not* believe that state-mandated and commercial large-scale assessments currently provide sufficient information about students' strengths and weaknesses. White bars represent the percentage of teachers who disagree. Gray bars represent the percentage of teachers who strongly disagree.

provided by large-scale assessments at the individual student and classroom levels, respectively; 62%, 56%, and 49% of teachers believed it would be valuable to have more diagnostic information available at the grade, school, and district levels, respectively (see Figure 2.10). Based on these results, demand for more diagnostic information from large-scale assessments appears to be strong, particularly at the individual student and classroom levels.

What Kinds of Diagnostic Information Do Teachers Want from Large-Scale Assessments?

In our survey of teachers, it was clear that most teachers wanted not only *more* diagnostic information than is typically provided on large-scale

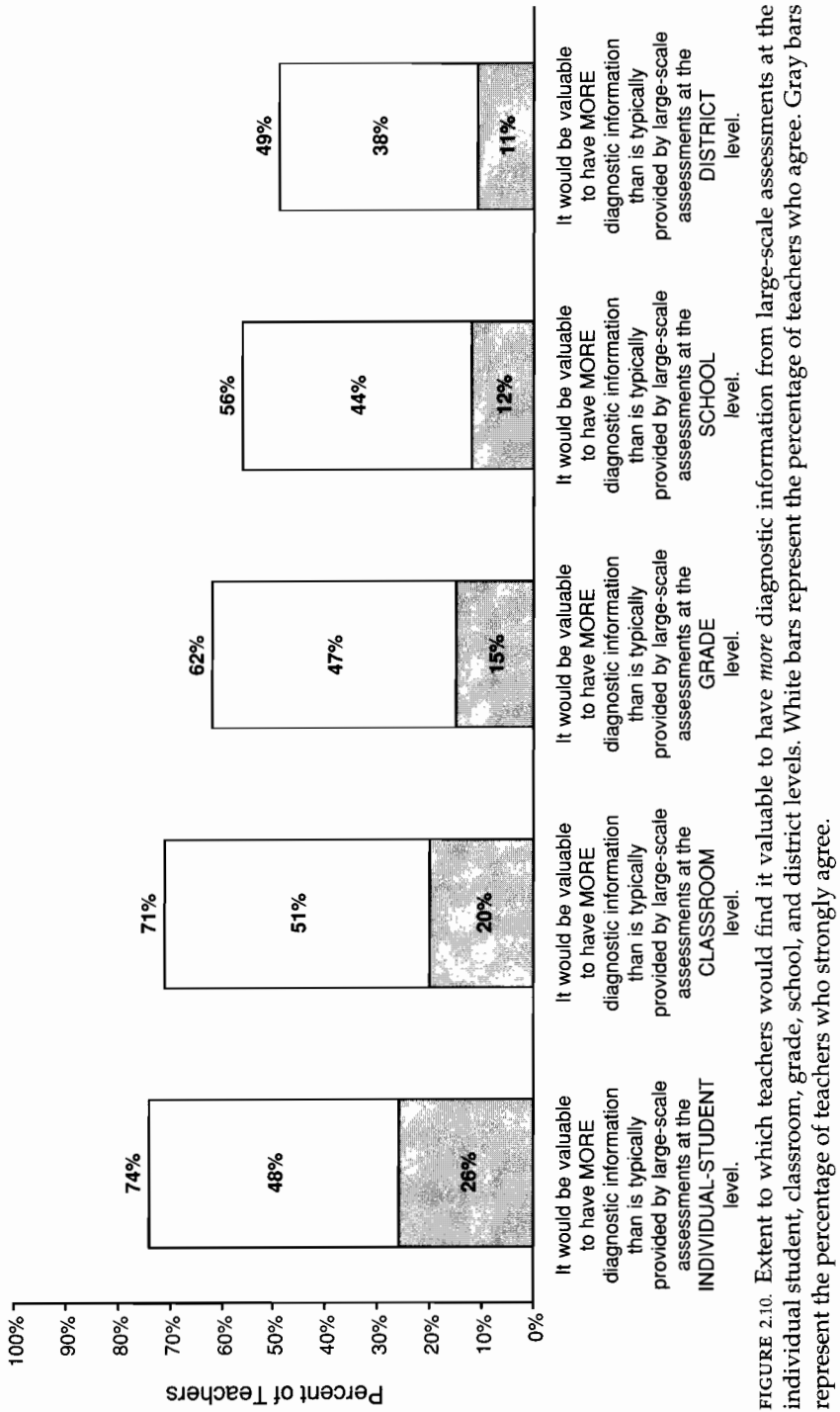

FIGURE 2.10. Extent to which teachers would find it valuable to have *more* diagnostic information from large-scale assessments at the individual student, classroom, grade, school, and district levels. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

assessments, but also *different* kinds of diagnostic information. As shown in Figure 2.11, 79% and 73% of teachers believed that it would be valuable to have different diagnostic information than is typically provided by large-scale assessments at the individual student and classroom levels, respectively. A majority of teachers (68%, 61%, and 56%) also believed that it would be valuable to have different diagnostic information at the grade, school, and district levels, respectively (see Figure 2.11).

A number of results from our survey provide insight into the types of diagnostic information that teachers want from large-scale assessments. As indicated in Figures 2.7C and 2.8C, approximately one-third of teachers (38% for state-mandated assessments and 30% for commercial assessments) believed that the diagnostic information currently reported for large-scale assessments is not detailed enough. Much smaller percentages (9% for state-mandated assessments and 11% for commercial assessments) held the contrary view that existing diagnostic information is too detailed (see Figures 2.7B and 2.8B).

We also asked teachers to indicate the importance of reporting assessment results using various points of reference. Although approximately one-half (51%) of teachers indicated that it is important or very important to report norm-referenced diagnostic information (see Figure 2.12I), a larger majority of teachers (between 64% and 89%) believed that it is important or very important to report diagnostic information using a variety of criterion-referenced approaches (see Figures 2.12A to 2.12H). Teachers appear to be especially interested in receiving descriptions of specific skills or knowledge individual students demonstrated on a large-scale assessment, as well as descriptions of specific skills or knowledge individual students should develop (with 89% and 85%, respectively, of teachers stating that it is important or very important to report diagnostic information in these ways) (see Figures 2.12A and 2.12B). Similarly, a majority of teachers indicated that it was important or very important to receive suggested strategies individual students might use to improve their skills or knowledge (81%), as well as to receive suggested strategies a teacher might use to address student needs (83%). Eighty percent of teachers believed that it was important or very important to report results in multiple ways. Seventy-four percent of teachers indicated that it was important or very important to report students' performance on individual items, and 68% and 64%, respectively, had similar beliefs for reporting diagnostic information using number and percentage of items correct (i.e., subscores) and standards-based information (i.e., results in relation to state performance standards).

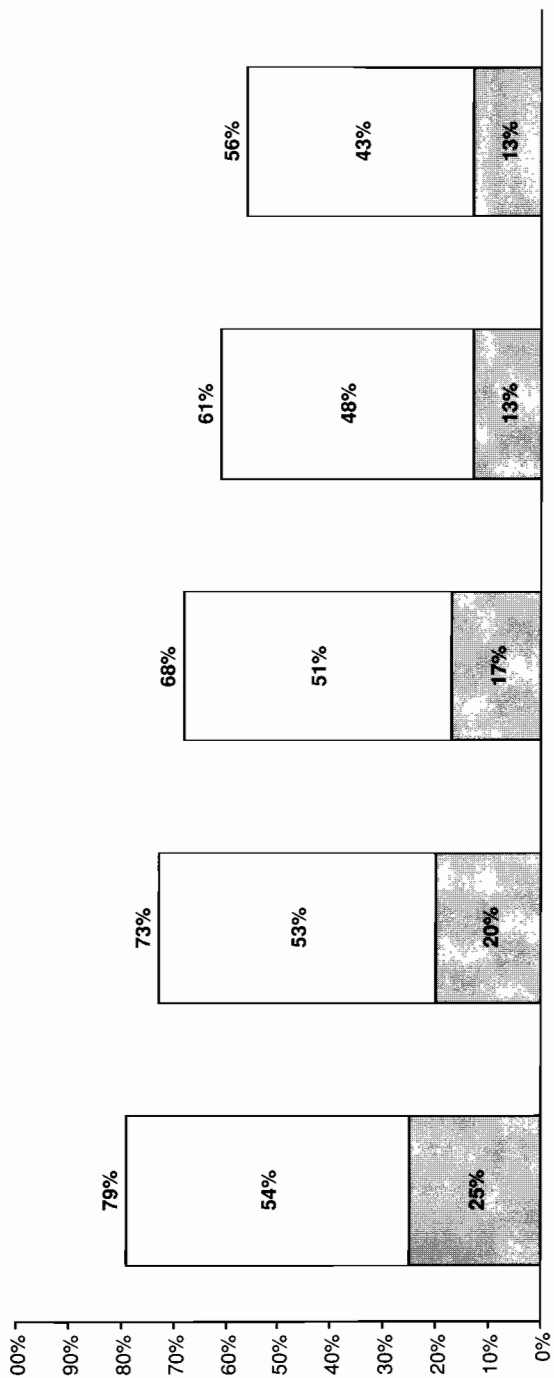

FIGURE 2.11. Extent to which teachers would find it valuable to have *different* diagnostic information from large-scale assessments at the individual student, classroom, grade, school, and district levels. White bars represent the percentage of teachers who agree. Gray bars represent the percentage of teachers who strongly agree.

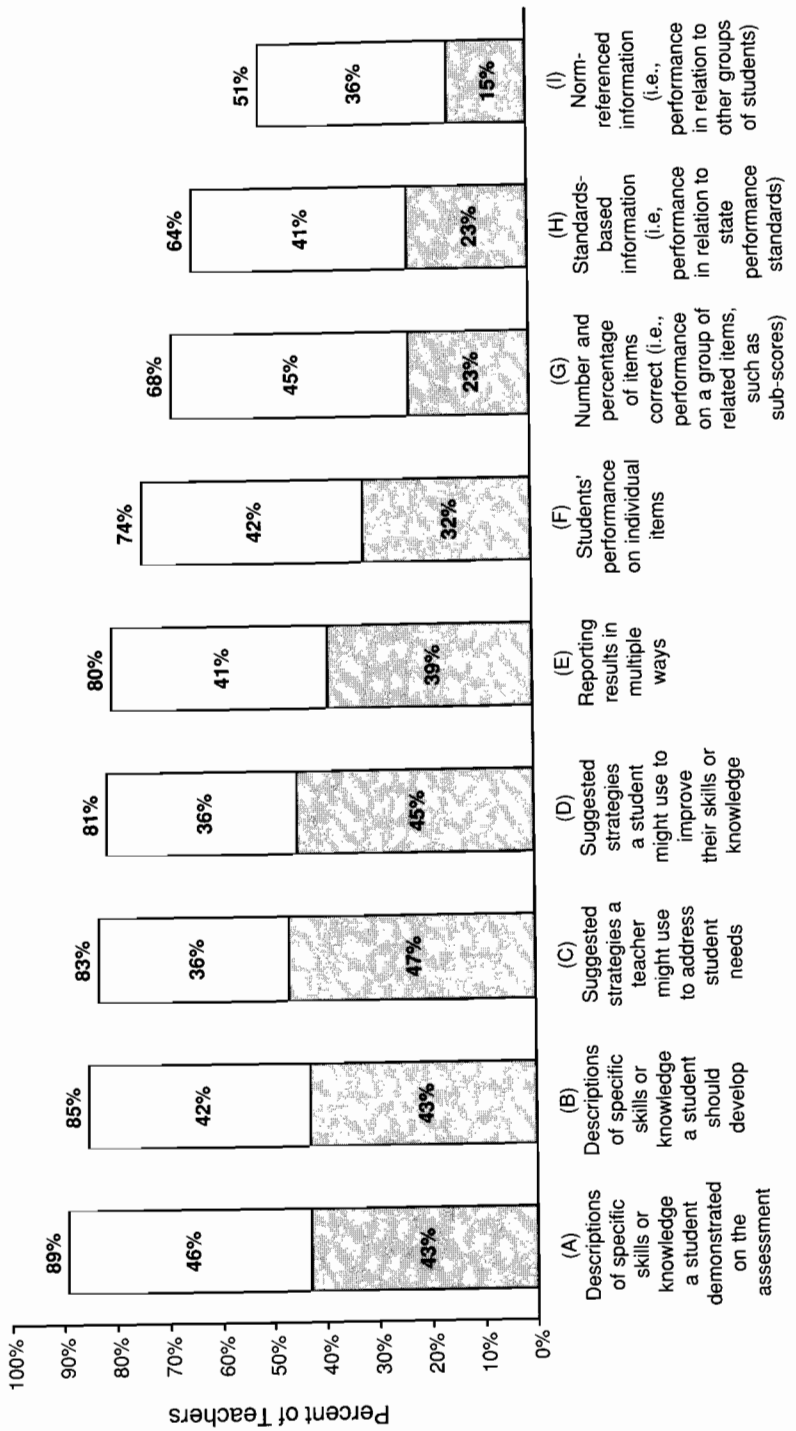

FIGURE 2.12. Extent to which teachers find various points of reference important for reporting diagnostic information. White bars represent the percentage of teachers who find various points of reference important. Gray bars represent the percentage of teachers who find various points of reference very important.

The high levels of importance that teachers give to these various methods of reporting diagnostic information, as well as research that indicates many of these reporting methods are not typically used in existing assessment programs (see Goodman & Hambleton, 2004), offer some important insight into ways that current reporting practices could be improved.

Summary of Educators' Demand for Diagnostic Information

The results of our survey demonstrate that educators are demanding both more and different diagnostic information from large-scale assessments. Although teachers are currently using the results of these assessments, there is a clear need to improve on existing diagnostic assessment reporting practices. In recent years, assessment developers have been modifying their practices to help address these needs. Some of these efforts are explored next.

SOME CURRENT EFFORTS TO ADDRESS EDUCATORS' NEEDS

Although educators appear to have become accustomed to receiving and using large-scale assessments results in their classrooms, the results of our survey suggest further efforts to provide improved and more relevant diagnostic information from these assessments is warranted. Assessment developers are currently responding to these issues in at least two ways: (a) by finding ways to better integrate large-scale assessment results into classroom practice, and (b) by using CDA principles and practices to improve diagnostic reporting procedures for existing assessments. Although neither approach reflects a full consideration of the principles that underlie CDA, we believe that they represent some promising ways to help address educators' needs and, especially if applied jointly, to implement some key CDA principles in practice.

Efforts to Help Integrate Diagnostic Assessment Results into Classroom Practice

In recent years, we have seen some promising efforts to better integrate diagnostic assessment results into classroom practice. These include the development of (a) classroom-based assessments that are directly related to state standards, and (b) guides for classroom and instructional planning that are aligned to information in the score reports of large-scale assessments.

An example of how diagnostic assessment results can be better integrated into classroom practice comes from Ohio. The Ohio Department of Education requires diagnostic testing of all first- and second-grade students in reading, writing, and mathematics. These assessments are designed to give teachers information about students' progress toward the state's standards (Ohio Department of Education, 2005). The assessments are designed to be flexible and thus vary in how they can be administered (e.g., assessments can be integrated into regular classroom activities, administered individually or to a group, and/or administered orally or written). For each state standard measured, students are classified into one of three categories – clearly on track, further assessment may be needed, or needs further assessment and/or intervention – based on the number of items they answered (Ohio Department of Education, 2005). Teachers are provided with recommended activities that are tailored to the needs of students who fall within a given category.

The Stanford 10 Achievement Test (SAT 10) provides another example of how diagnostic assessment results can be better integrated into classroom practice. The SAT 10 is a commercial large-scale assessment that reports diagnostic information in the form of general subscores (e.g., Word Study Skills, Reading Vocabulary, Reading Comprehension on the Grade 4 Reading test) and more detailed subcategories of each subscore (e.g., Literary, Informational, Functional, Initial Understanding, Interpretation, Critical Analysis, Strategies, Thinking Skills; Harcourt Assessment, 2006a). A lot of data (e.g., number correct, scale score, national percentile rank, national grade percentile, grade equivalent) are provided for each subscore. Further information (number of items, number answered, and number correct) is provided for the more detailed subcategories of each subscore. Given the large amounts of data that are provided, there is a clear risk that teachers may find the information difficult to properly interpret and use (a position that is informed by the results of our survey, as well as by research of Impara, Divine, Bruce, Liverman, and Gay, 1991). However, the availability of resources that go beyond traditional interpretive guides should help in this regard. Along with the score reports, educators can also receive guides for classroom and instructional planning that are aligned to the information in the score report (Harcourt Assessment, 2006b). Although our research did not specifically address the efficacy of these specific supplementary materials, it is clear from our results that teachers are currently lacking resources to inform the proper use of diagnostic information from large-scale assessments. Based on this finding, we believe that any responsible

attempts to develop resources that link classroom practices with diagnostic assessment results is a step in the right direction and should be explored by all assessment developers.

Using Cognitive Diagnostic Assessment Principles and Practices to Improve Diagnostic Reporting Procedures for Existing Large-Scale Assessments

The number of assessment developers who discuss, research, and implement principles and practices that have been informed by the CDA literature is growing. Nichols (1994) and NRC (2001) provide several examples of diagnostic assessment systems designed and built completely within a cognitive framework (e.g., HYDRIVE, GATES, Automated Cognitive Modeler). However, because most large-scale assessments have been developed through a psychometric approach to assessment design and often were not originally designed to provide diagnostic feedback, it is helpful to consider what steps, if any, can be taken to improve the instructional relevance of the results from these assessments when the full implementation of CDA principles is not feasible. We explore this issue by examining some ways in which CDA principles have been applied to improve the scoring and reporting procedures of existing psychometrically designed assessments.

Testing programs of all types are exploring alternative approaches to scoring and reporting that will allow them to report sufficiently reliable information at smaller and smaller grain sizes in the hope that such feedback can inform instruction. Methodologies that are frequently researched include augmenting subscores with Bayesian techniques (Wainer et al., 2001) or multidimensional item response theory (Thissen & Edwards, 2005), as well as comparing individual item response patterns to item-by-skill matrices to classify students as masters or nonmasters of each skill, such as with rule space methodology (Tatsuoka, 1983, 1995). These approaches are a departure from the typical methods of providing diagnostic information using number and percent correct or performance on individual items, and are of interest as assessment developers seek to improve current practice.

CDA principles are also influencing testing programs that were not designed to be diagnostic by enabling them to provide test candidates with more than just a scale score and the associated percentile rank, and to identify different skills for which they can report results. Some recent efforts by the College Board help illustrate this fact.

The PSAT/NMSQT Score Report Plus, which was first provided to students and educators in 2001, employed a cognitive-centered approach in its development. Researchers and test developers identified the multiple skills required to solve each item (Buck et al., 1998), and then used a modified rule-space approach (DiBello, 2002; DiBello & Crone, 2001) to classify students as either masters or nonmasters on each of the skills according to their individual item response patterns. Suggestions for how to improve on weak skills are also provided on the score report. The PSAT/NMSQT was not defined from an explicit cognitive model, but this CDA-influenced approach to providing individualized score reports was seen as a more effective way to provide instructionally useful information than approaches that do not differentiate strengths and weaknesses among students with the same total score, or subscores grouped by content or skill.

A second example of CDA influence on large-scale testing programs is the College Board research initiative to identify cognitive skills for the critical reading and mathematics sections of the SAT in an effort to make SAT results more informative to both students and educators. In 2003, as the development for the revised SAT was underway, research began on how to better describe student performance in ways that could help inform instruction. At the time, the current SAT score report only provided raw number correct and estimated percentile scores for items grouped by item type on the Verbal section (i.e., Critical Reading, Analogies, and Sentence Completion) and items grouped by content category on the Mathematics section (i.e., Arithmetic and Algebraic Reasoning, and Geometric Reasoning). The value of this type of feedback was questionable, especially when one tries to imagine how such feedback could inform instruction. Although the SAT was not designed within an explicit model of cognition in reading or mathematics, in an effort to improve the diagnostic utility of the test results, the College Board initiated research to determine the types of cognitive skills that underlay the test (Huff, 2004; O'Callaghan, Morley, & Schwartz, 2004; VanderVeen, 2004). As a result of this research, the College Board is able to demonstrate that the critical reading section can be theoretically and empirically supported as measuring the following clusters of text comprehension skills: determining the meaning of words; understanding the content, form, and functioning of sentences; understanding the situation implied by a text; understanding the content, form, and function of larger sections of text; and analyzing authors' purpose, goals, and strategies (VanderVeen et al., 2007). As noted by VanderVeen et al.,

these five clusters generally align with extensively researched cognitive processing models of text comprehension, such as the ones proposed by Kintsch (1998) and Perfetti (1985, 1986), providing evidence to support the validity of these potential new reporting classifications and demonstrating that the test as a whole assesses the types of cognitive processes that underlie text comprehension. Thus, in addition to providing additional skill areas for which results can be reported, this type of post hoc analysis offers useful ways for test developers to show that their assessments align with the cognitive models that underlie the construct of interest (a powerful piece of evidence for establishing test validity).

These examples of improving the instructional relevance of large-scale assessments are encouraging. In the next section, we discuss future directions for those interested in improving the instructional utility of assessment results.

FUTURE DIRECTIONS

In this chapter, we explore the demand that exists for CDA from within the assessment community and educators' demands for diagnostic feedback from the assessments for which their students regularly take part. We have also outlined some promising ways to make diagnostic results more relevant to instructional practice, as well as some efforts to improve the types of diagnostic information that can be derived from existing large-scale assessments. We end this chapter by highlighting some issues that, based on the demands discussed here, warrant further attention. These include efforts to develop more coherent, comprehensive, and continuous assessment systems that are based on the same underlying model of learning, and take advantage of recent technological advancements. Also, to maximize potential, assessment results must be provided to educators in a timelier fashion, and the assessment results must be presented in a more useful and meaningful manner.

Develop Assessment Systems Based on the Same Underlying Model of Learning

The intent of providing detailed diagnostic feedback from large-scale assessments is to facilitate valid interpretations with regard to students' strengths and weaknesses on the material tested, as well as to aid in generalizing from the assessment context to the domain of interest. Presumably, the value of assessment results increase as their interpretability

in the context of classroom instruction and learning increases. As mentioned previously in the chapter, when curriculum, instruction, and assessment are part of a coherent system that is based on a common model of learning, then the assessment results, per force, will be instructionally relevant. The application of CDA principles in the design and development of assessments is certain to help in this regard, and should be the ultimate goal of assessment developers. However, in situations where this is not immediately feasible (e.g., in the case of well-established assessment programs), we propose that adapting CDA principles in the analyses and reporting of results (e.g., what has been done for the PSAT/NMSQT and SAT) is a reasonable and practicable response to address some important demands from assessment developers who advocate CDA and from educators who are demanding better diagnostic information from large-scale assessments.

Although the primary focus of this chapter was to explore the demands for CDA from the perspective of large-scale testing, it is clear from the results of our survey that teachers overwhelmingly view classroom-based assessments as being the best way to gauge student strengths and weaknesses. Unfortunately, however, a body of research exists that shows that classroom assessment practices do not always provide accurate and valid information, measure a full complement of cognitive skills (see, e.g., Notar et al., 2004, and Stiggins, 2001), or are as well integrated with instruction as they could be (Black & Wiliam, 1998a, 1998b). Consequently, a more concerted effort by the measurement community and teacher training programs to provide teachers with a full array of assessment tools (particularly those that serve a formative function) appears warranted. Many have discussed the benefits of using a cognitive model of learning as the basis for an integrated system of curriculum, instruction, and assessment in the classroom (e.g., Nichols, 1993, 1994; NRC, 2001; Pellegrino et al., 1999). Consistent with the views of NRC (2001), we argue that extending the application of CDA principles to the development of large-scale assessments would enable the creation of a comprehensive assessment system, where results from all levels would provide complementary pieces of evidence that can be readily linked to and interpreted within a common model of student learning.

Take Advantage of Recent and Upcoming Technological Advancements

As noted by NRC (2001), technological advancements are helping remove some of the constraints that have limited assessment practice

in the past. Assessments no longer need to be confined to a paper-and-pencil format. Computer-based platforms will help assessment developers use innovative item types that have the potential to measure the kinds of knowledge and skills that are more reflective of the cognitive models of learning on which assessments should be based (Huff & Sireci, 2001; Sireci & Zenisky, 2006). The implementation of computer-based assessment systems will also enable assessment developers to move away from traditional scoring methods that only consider whether a student provided the correct response to an item by allowing them to collect data on such things as the choices that students make in items that have multiple components, auxiliary information accessed when answering an item, and the length of time students take to complete an item (Luecht, 2002). Of course, these advancements also introduce a number of technical challenges outside the more general challenges (e.g., cost and the current lack of suitable equipment in many K–12 schools) that currently limit the widespread application of computer-based assessments in the K–12 environment. These include developing complex scoring models that can make use of all data that are available from computer-based assessments to make valid inferences about the cognitive strategies employed by the examinee, to diagnose the learning state of the examinee, and to provide instructionally relevant feedback that can be made available to teachers and students in a time frame that will increase the use and value of these results. Furthermore, it is hoped that other creative solutions that exploit technological advances, such as automated scoring that minimizes or removes the need for human scorers, can be refined and used within the operational constraints of K–12 large-scale assessments. Thus, although more general issues such as cost and the limited availability of necessary equipment in the K–12 system are being addressed by policy makers, researchers should ready themselves for the day when computers are accessible to each examinee and should investigate ways to use technology to assess new and important skills of interest, such as the ability to transfer knowledge to new situations.

Improve Operational Procedures to Minimize the Delay in Reporting Results

In our survey, we found that a key factor that inhibits teachers' use of diagnostic information from large-scale assessments is the significant delay between test administration and the availability of results. We see this as major obstacle for effective use of large-scale assessment data, and

one that deserves greater attention if these assessments are to provide useful diagnostic information to educators, parents, and students.

The need to minimize lag time between administering the test and reporting results is something of which most assessment developers and policy makers are well aware. Still, it is the unfortunate reality that many large-scale assessment results are released months after the tests are administered, and often are not released until the subsequent school year (after students have moved on to another teacher and possibly to another school). If assessment developers want to provide educators with meaningful and useful diagnostic information, they must find ways to do this in a timelier manner, without sacrificing the quality and integrity of the assessments. It is encouraging to note that some departments of education have made significant progress in this area. For example, the British Columbia (BC) Ministry of Education, which serves approximately 575,000 K–12 public school students (BC Ministry of Education, 2005a, p. 11), has refined its assessment cycle and administration and reporting procedures so they are able to release results only *4 weeks* after an assessment session (BC Ministry of Education, 2005b, pp. 33–34). In addition to offering multiple testing sessions throughout the year, the BC Ministry of Education (2005b, 2006) delivers student-, school-, and district-level results electronically through secure Internet portals, and now provides schools with the option of administering a number of large-scale assessments on computers. We are encouraged by these developments, and it is hoped that other K–12 assessment programs will effectively improve the speed with which they are able to release assessment results, while maintaining (and even improving on) the quality of their assessments.

Present Assessment Results in a More Useful and Meaningful Manner

Many techniques can be applied to improve the types and clarity of information that are provided on large-scale assessment score reports. In their review of existing student-level score reports from state-mandated and commercial assessments, Goodman and Hambleton (2004) outlined some weaknesses in current reporting practices that should be addressed. These include reporting an excessive amount of information in some reports and not reporting other essential pieces of information (e.g., the purpose of the assessment and how the results will and should be used) in others, not providing information about the precision of the test scores, using statistical jargon that will not be readily understood by users of the reports, and reporting a large amount of information in too small of a space. Further guidance is also available in the work of Hambleton and Slater (1997), Impara et al. (1991), Jaeger (1998), Wainer (1997), and Wainer, Hambleton, and Meara (1999).

The research presented here shows that teachers are eager for both more and different types of diagnostic information. Suggestions have been provided on how this type of information could be made available using CDA principles. Presenting this information in a useful and meaningful manner is another challenge that assessment developers must face and address.

CONCLUSION

Large-scale assessments for summative and high-stakes purposes are an integral part of the K–12 educational system. Although these assessments are typically developed for the purposes of accountability or to rank-order students, redesigning these assessments from a cognitively principled approach could help integrate these assessments with teaching and learning in the classroom without necessarily jeopardizing their primary purposes. Similarly, CDA principles and practices can be used post hoc to identify new types of information that can be reported for assessments developed from within a psychometric framework. Results from our survey show that teachers are searching for as much information as possible about their students from various sources. It is our responsibility as educators to respond creatively to their needs by taking advantage of new knowledge about instruction and learning when designing assessments and when analyzing and reporting assessment results. The application of CDA principles would be an important advancement in this regard.

References

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*(3), 5–12.

Bennett, R. E., & Bejar, I. I. (1998).Validity and automated scoring: It's not only the scoring. *Educational Measurement, 4*, 9–17.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*(2), 62–176.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

British Columbia (BC) Ministry of Education. (2005a). *2004/05 Service plan report*. Retrieved June 26, 2006, from http://www.bcbudget.gov.bc.ca/Annual_Reports/2004_2005/educ/educ.pdf.

British Columbia (BC) Ministry of Education. (2005b). *Handbook of procedures for the graduation program*. Retrieved June 26, 2006, from http://www.bced.gov.bc.ca/exams/handbook/handbook_procedures.pdf.

British Columbia (BC) Ministry of Education. (2006). *E-assessment: Grade 10 and 11 – Administration*. Retrieved June 26, 2006, from http://www.bced.gov.bc.ca/eassessment/gradprog.htm.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal: Sentence completion section* (ETS Research Report [RR-98–23]). Princeton, NJ: Educational Testing Service.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 1–18). Mahwah, NJ: Erlbaum.

College Board. (2006). *Passage-based reading*. Retrieved January 15, 2006, from http://www.collegeboard.com/student/testing/sat/prep_one/passage_based/pracStart.html.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (3rd ed., pp. 443–507). Washington, D.C.: American Council on Education.

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

DiBello, L. V. (2002, April). *Skills-based scoring models for the PSAT/NMSQT™*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

DiBello, L. V., & Crone, C. (2001, April). *Technical methods underlying the PSAT/NMSQT™ enhanced score report*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Erlbaum.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407–433.

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1993). Learning, teaching, and testing for complex conceptual understanding. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 181–218). Hillsdale, NJ: Erlbaum.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6*(4), 397–416.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220.

Goodman, D. P., & Huff, K. (2006). *Findings from a national survey of teachers on the demand for and use of diagnostic information from large-scale assessments*. Manuscript in preparation, College Board, New York.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generations. *Journal of Educational Measurement, 42*, 351–373.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles: National Center for Research on Evaluation, Standards, and Student Teaching.

Harcourt Assessment. (2006a). *Stanford achievement test series, tenth edition: Critical, action-oriented information*. Retrieved January 29, 2006, from http://harcourtassessment.com/haiweb/Cultures/en-US/dotCom/Stanford10.com/Subpages/Stanford+10+-+Sample+Reports.htm.

Harcourt Assessment. (2006b). *Support materials for parents, students, and educators*. Retrieved January 29, 2006, from http://harcourtassessment.com/haiweb/Cultures/en-US/dotCom/Stanford10.com/Subpages/Stanford+10+-+Support+Materials.htm.

Huff, K. (2004, April). A practical application of evidence centered design principles: Coding items for skills. In K. Huff (Organizer), *Connecting curriculum and assessment through meaningful score reports*. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(4), 16–25.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers?*Educational Measurement: Issues and Practice, 10*(4), 16–18.

Jaeger, R. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*.Cambridge, MA: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Luecht, R. M. (2002, April). *From design to delivery: Engineering the mass production of complex performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Massachusetts Department of Education. (2004). *2004 MCAS technical report*. Retrieved January 15, 2006, from http://www.doe.mass.edu/mcas/2005/news/04techrpt.doc#_Toc123531775.

Massachusetts Department of Education. (2005). *The Massachusetts comprehensive assessment system: Guide to the 2005 MCAS for parents/guardians*. Malden: Author.

Mathematical Sciences Education Board. (1993). *Measuring what counts: A conceptual guide for mathematics assessment.*Washington, DC: National Academy Press.

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*(4), 379–416.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International and University of Maryland. Retrieved May 1, 2006, from http://padi.sri.com/downloads/TR9_ECD.pdf.

Missouri Department of Elementary and Secondary Education. (2005). *Missouri assessment program: Guide to interpreting results.* Retrieved June 24, 2006, from http://dese.mo.gov/divimprove/assess/GIR_2005.pdf.

National Research Council (NRC). (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

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

National Research Council (NRC). (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools.*Washington, DC: National Academy Press.

New Jersey Department of Education. (2006). *Directory of test specifications and sample items for ESPA, GEPA and HSPA in language arts literacy.* Retrieved June 24, 2006, from http://www.njpep.org/assessment/TestSpecs/LangArts/TOC.html.

Nichols, P. D. (1993). *A framework for developing assessments that aid instructional decisions* (ACT Research Report 93–1). Iowa City, IA: American College Testing.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575–603.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447–474). New York: American Council on Education/Macmillan.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 1111, 115 Stat. 1449–1452 (2002).

Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology, 31*(2), 115–129.

O'Callaghan, R., Morley, M., & Schwartz, A. (2004, April). *Developing skill categories for the SAT® math section.* In K. Huff (Organizer), Connecting curriculum and assessment through meaningful score reports. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

Ohio Department of Education. (2005). *Diagnostic guidelines.* Retrieved February 1, 2006, from http://www.ode.state.oh.us/proficiency/diagnostic_achievement/Diagnostics_PDFs/Diagnostic_Guidelines_9-05.pdf.

O'Neil, T., Sireci, S. G., & Huff, K. L. (2004). Evaluating the content validity of a state-mandated science assessment across two successive administrations of a state-mandated science assessment. *Educational Assessment and Evaluation, 9*(3–4), 129–151.

Pellegrino, J. W. (2002). Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment. In M. J. Smith (Ed.), *NSF K–12 Mathematics and science curriculum and implementation centers conference proceedings* (pp. 76–92). Washington, DC: National Science Foundation and American Geological Institute.

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.

Perfetti, C. A. (1985). Reading ability. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 31–58). New York: W. H. Freeman.

Perfetti, C. A. (1986). *Reading ability.*New York: Oxford University Press.

Riconscente, M. M., Mislevy, R. J., & Hamel, L. (2005). *An introduction to PADI task templates* (PADI Technical Report 3). Menlo Park, CA: SRI International and University of Maryland. Retrieved May 1, 2006, from http://padi.sri.com/downloads/TR3_Templates.pdf.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333–352.

Sheehan, K. M., Ginther, A., & Schedl, M. (1999). *Development of a proficiency scale for the TOEFL reading comprehension section* (Unpublished ETS Research Report). Princeton, NJ: Educational Testing Service.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–348). Mahwah, NJ: Erlbaum.

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.

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., DiBello, L. V., Senturk, D., Yan, D., Chernick, H., & Kindfield, A. C. H. (2003). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability* (CRESST Technical Report 609). Los Angeles: Center for the Study of Evaluation, CRESST, UCLA.

Stiggins, R. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*(3), 5–15.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Thissen, D., & Edwards, M. C. (2005, April). *Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC*

strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

VanderVeen, A. (2004, April). *Toward a construct of critical reading for the new SAT*. In K. Huff (Organizer), Connecting curriculum and assessment through meaningful score reports. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies*. Mahwah, NJ: Erlbaum.

Wainer, H. (1997). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, *22*(1), 1–30.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*(4), 301–335.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented scores: "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Erlbaum.

Washington State Office of Superintendent of Public Instruction. (2006). *Test and item specifications for grades 3–high school reading WASL*. Retrieved June 24, 2006, from http://www.k12.wa.us/Assessment/WASL/Readingtestspecs/TestandItemSpecsv2006.pdf.

3

Cognitive Modeling of Performance on Diagnostic Achievement Tests

A Philosophical Analysis and Justification

Stephen P. Norris, John S. Macnab, and Linda M. Phillips

To interpret and use achievement test scores for cognitive diagnostic assessment, an explanation of student performance is required. If performance is to be explained, then reference must be made to its causes in terms of students' understanding. Cognitive models are suited, at least in part, to providing such explanations. In the broadest sense, cognitive models should explain achievement test performance by providing insight into whether it is students' understanding (or lack of it) or something else that is the primary cause of their performance. Nevertheless, cognitive models are, in principle, incomplete explanations of achievement test performance. In addition to cognitive models, normative models are required to distinguish achievement from lack of it.

The foregoing paragraph sets the stage for this chapter by making a series of claims for which we provide philosophical analysis and justification. First, we describe the philosophical standpoint from which the desire arises for explanations of student test performance in terms of causes. In doing this, we trace the long-held stance within the testing movement that is contrary to this desire and argue that it has serious weaknesses. Second, we address the difficult connection between understanding and causation. Understanding as a causal factor in human behavior presents a metaphysical puzzle: How is it possible for understanding to cause something else to occur? It is also a puzzle how understanding can be caused. We argue that understanding, indeed, can cause and be caused, although our analysis and argument are seriously compressed for this chapter. Also, in the second section, we show why understanding must be taken as the causal underpinning of achievement tests. Third, we examine how cognitive