



Commentary on the National Mathematics Advisory Panel Recommendations on Assessment

Lorrie A. Shepard

Foundations for Success: The Final Report of the National Mathematics Advisory Panel (2008) recommends that National Assessment of Educational Progress and state tests focus on foundations of algebra and include fewer pattern-type algebra items. Greater mathematics expertise is needed in test development to prevent flawed items. The Panel reached the erroneous conclusion that multiple-choice and constructed response items measure the same thing because it relied on studies where the two item types were constrained to be identical. The Panel's recommendations on standard setting are misleading because it relied on a single simulation study. By focusing only on special education studies, its discussion of formative assessment implies that formative assessment requires formal tools administered and scored frequently and fails to recognize more interactive forms of feedback found in studies from cognitive science.

Keywords: assessment; NAEP; standard setting

Assessment is clearly one of the most important aspects of education policy and practice and one that is more affected by federal policy than even instruction and curriculum. The National Mathematics Advisory Panel (NMAP; 2008) chose to focus on the National Assessment of Educational Progress (NAEP) and six state tests. The Panel brought its content expertise to bear in examining released items from the test and in reviewing test blueprints. As a result of its findings, the Panel argued for better representation of mathematicians “along with mathematics educators, mathematics education researchers, curriculum specialists, classroom teachers, and the general public in the standard-setting process and in the review and design of mathematical test items for state, NAEP, and commercial tests” (p. 60). The Panel is at its best, even when discussing assessment (chapter 9, “Assessment of Mathematics Learning”), when talking about mathematics content, and it brings to bear both research evidence and members’ professional opinions. Oddly, however, despite the participation of very strong measurement experts, there was not a clear conceptual framing of measurement issues, and only one noncontent recommendation made it all the way to the executive summary.¹ It says essentially that NAEP and

states should develop procedures for item development that ensure the highest mathematical and psychometric quality. Yes, of course.

This commentary is organized around the main measurement issues considered by the Panel: performance categories and test and item design. I also discuss the Panel's treatment of formative assessment in chapter 7 (“Instructional Practices”). I do not provide a detailed review of its assessment content recommendations, which are acknowledged here only briefly. There is ample evidence that the unintentional U.S. mathematics curriculum is too broad, chaotic, and redundant and needs to be made more coherent and focused (Schmidt, McKnight, & Raizen, 1997). This means correspondingly that a refocusing of assessment content must also be considered. Other reviewers have addressed the Panel's specific recommendations regarding the Critical Foundations of Algebra. The Panel's recommendation to expand the Number Properties and Operations component of the NAEP is buttressed by findings from the more detailed *Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8* (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007), funded by the National Center for Education Statistics.

On the smaller issue of too many pattern items being used to measure algebra, of course this needs to be corrected. (Pattern items are typically in the form of a number series or series of figures, and students are asked to pick which number or figure will come next.) Narrow representation of a construct by overusing specific item types is a major threat to the validity of test score inferences and is the primary cause of negative teaching-the-test effects and the failure of test scores to generalize and learning to transfer. Also, both the Daro et al. (2007) study and the Panel's own review of state and NAEP assessments found an unacceptably large number of flawed items, such that aspects of the item format or wording would possibly interfere with valid assessment of the intended mathematical knowledge or skill. As suggested by the Panel, greater mathematical expertise is needed for both item development and review—a recommendation, I will note, that is harder and harder to implement as the amount of testing increases. Regarding the use of calculators on assessments, a reasonable summary, both logically and empirically, would be that calculators should not be used when assessing students' computational skills, but when the focus of assessment is problem-solving ability, calculators can improve performance without harming validity.

Although the Panel picked a few important measurement topics to consider, there are also crucial issues left out that should

Educational Researcher, Vol. 37, No. 9, pp. 602–609
DOI: 10.3102/0013189X08328001
© 2008 AERA. <http://er.aera.net>

be mentioned before moving on. For example, the Panel did not say much about the dilemmas inherent in creating a national assessment in a country without a national curriculum. Commercial test publishers have tended to solve this problem by developing “lowest common denominator,” basic-skills tests. Think of these as the intersection of all possible curricula. In contrast, NAEP has historically taken the stance of providing a “comprehensive” domain framework that is essentially the union of all possible curricula. At least this avoids the dumbing-down effects of testing only the minimums, but does this exacerbate the mile-wide and an inch-deep phenomena? Rather than merely proceeding to make its own content recommendations, it would have been helpful for the Panel to consider both the logistics (how similar are existing state frameworks to one another) and the politics of arriving at a more focused and coherent set of assessment frameworks. The Panel also missed an opportunity by not citing *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), a landmark National Research Council report on assessment and the critical need it identified for coherence between large-scale and classroom-level assessments.

Perhaps most seriously, the Panel failed to consider the research evidence on teaching-the-test effects and, therefore, was unable to make recommendations about how assessment design can foster or forestall the lack of learning generalization that occurs when teaching looks too much like practice on the test. In the introduction, the Task Group on Assessment report says, “Tests make visible to teachers, parents, and policymakers some of the outcomes of student learning. They also can drive instruction” (NMAP, 2008, p. 57). There is no further discussion of this important idea. The Panel implies that the issue is merely a matter of choice. If algebra is important, include more on the test and students will learn more of it. The Panel does not acknowledge the distortions that can occur under a test-driven curriculum, whereby test scores go up and students appear to be mastering content when in fact they are not.

Herman (2008) and Shepard (2008) provide up-to-date summaries of the well-known literature on the effects of high-stakes testing. Testing systematically redirects teaching effort. Of course, the logic model underlying standards-based reform intends that instruction and assessment should be mutually aligned with standards, and in some cases the evidence suggests that indeed attention to the test has moved instruction in positive directions intended by reform. In Washington, for example, Stecher and Chun (2001) found that teachers reported spending more time on probability and statistics and on “sense-making” activities such as representing and sharing information, relating concepts, and formulating questions, directly in response to the state learning goals and new assessments. In a great many cases, however, teaching-the-test instructional practices so closely resemble the test that it is unlikely that students have the opportunity to gain deeper conceptual understanding. When teacher survey data and observational studies are combined with large-scale studies that test the generalizability of test score gains (see Koretz, 2008, for a review), it becomes apparent that many of the instructional strategies used to raise test scores do not have a commensurate positive effect on student learning. Koretz and Hamilton (2006) developed a typology to help analyze test preparation activities to evaluate whether

they are likely to lead to true gains in achievement, score inflation, or both. More recently, Koretz (2008) has called for better test design that anticipates the likelihood that teachers will imitate particularities of item formats and builds in greater variation across dimensions of generalization to prevent the types of distortions in teaching and learning that have been documented so extensively. In my view, these understandings of what it takes to protect a national monitoring assessment and to build adequate state accountability tests are as important as the Panel’s substantive recommendations about algebra and number.

Performance Categories

Performance standards, and especially state-to-state variation in the meaning of state standards, is one of the hottest issues in large-scale-assessment policy today. *Performance standards* (also called *proficiency standards*) refer to the cut points or cut scores on the total test score scale that separate performance categories, such as the advanced, proficient, basic, and below basic categories on NAEP. For some time, *Education Week* in particular has reported the startling inconsistencies among state proficiency rates and between results on state tests and NAEP (Olson, 2002, 2005). For example, on the 2000 NAEP Eighth-Grade Mathematics Assessment, only 30% of the students in North Carolina scored at proficient or above, in contrast to 81% who were said to be proficient on the state’s own test. In neighboring South Carolina, the state and national results were much more consistent: 20% proficient on the state test and 18% proficient on NAEP (Olson, 2002). Although these differences could be due to differences in the difficulty of the tests, to differences in how well aligned tests are with what is taught, or to sampling error, most measurement specialists have known that these fluctuations are primarily due to differences in the stringency of state performance standards themselves. Recently, a more technically elegant mapping study was done that verified, through statistical linkages (National Center for Education Statistics, 2007), that indeed quiet different and generally more lax standards have been set on state tests than on NAEP.

These large differences in the meaning of proficiency from state to state have led to considerable confusion among policy makers and the public about the interpretability of assessment results. (This confusion is especially ironic given that NAEP’s achievement levels were introduced in the first place to help make test results more interpretable to the public.) Although some of the differences in proficiency cutoffs can be attributed to the differences in the standard-setting methods employed (Linn, 2003), by far the greater cause of differences is the political context operating at the time the standards were set. In some cases, cut points were set to reflect minimum competencies, whereas in other cases standard setters were exhorted to aim for “world class” standards (Shepard, 2008). Although set at different times and for different purposes, state proficiency standards now have serious consequences for schools because they are the basis for determining adequate yearly progress under No Child Left Behind. Reporting by percentage proficient, instead of school averages, also poses other technical and policy problems. Student progress is ignored in parts of the score distribution away from the cut score, hence the phenomenon of focusing effort on the “bubble kids” (Hamilton et al.,

2007), and the apparent opening and closing of the achievement gap can be an artifact of the cut score location rather than real progress for minority groups as a whole (Holland, 2002). Perhaps it is time for policy makers who insisted on performance standards to know “how good is good enough” to reexamine whether reporting in proficiency metrics really does help with the interpretability of results. Unfortunately, statistical cutoffs set on a broad assortment of items, decoupled from any particular curriculum, do not have the intended meaning of substantive benchmarks set in the context of curriculum-based learning progressions.

Having identified performance categories as an important issue, it is disappointing and odd that the Panel elected to study a more manageable but significantly less important aspect of that problem. The Panel’s report focuses on the adequacy of the process, not on the validity of the result. It mentions in passing that the NAEP standards are too high in comparison to international data, but it does not address how external validity evidence should be brought into the standard-setting process. Perhaps without realizing it, the task group chose one side of the debate in the technical community. The task group repeatedly cites Reckase (2006) as its authority on standard-setting methods but ignores the chapter on performance standards in a significant National Research Council report (Pellegrino, Jones, & Mitchell, 1999) along with all of the previous studies investigating the NAEP achievement levels.

Reckase (2006) conducted a simulation study to evaluate two standard-setting methods. As he explained, there are two ways of thinking about the adequacy of methods for estimating a judge’s intended cut score. One uses a reliability approach and asks how small the standard errors are around an intended cut score, and the other takes a validity approach asking whether the intended score is supportable given external validity evidence. Reckase considered only the first approach and acknowledged that the validity of performance standards was beyond the scope of his study. It should also be noted that the items in Reckase’s simulation were constrained to be consistent with the Rasch model. Therefore, it is unlikely that the modified Angoff procedure was challenged in his study the way it has been with real data sets where hugely inconsistent results occurred for multiple-choice versus short-answer items and for right–wrong versus extended response items (Shepard, Glaser, Linn, & Bohrnstedt, 1993). Although Reckase’s article is a useful simulation study, it is by no means a comprehensive treatment of the corpse of work in this area and should not be the basis for making recommendations to the field.

It is a mistake to talk about standard-setting procedures as being “scientific” because this implies that by rigorous, systematic, and objective means expert judges can somehow get to the truth of the matter, thus discovering where the true cut score should be. It is more appropriate when the task group refers to the modified Angoff and Bookmark procedure as “professionally acceptable” procedures. Both the National Academy of Education study (Shepard et al., 1993) and the National Research Council report (Pellegrino et al., 1999) concluded,² however, that the modified Angoff procedure used to set achievement levels on NAEP was fundamentally flawed. They did so on the basis of the very large internal inconsistencies noted above.

Expert judges, even when they know a content area well, have a very difficult time translating substantive descriptions of standards into estimated passing rates on individual test items. As the task group noted, both NAEP and state assessments use very abstract, global definitions of desired knowledge and skills and therefore “would require high degrees of judgment to determine the categorization of student performance” (p. 8). Across studies there is a consistent tendency for expert judges to be able to order items by difficulty, roughly, but not to be able to tell at all how far apart they are on the score scale. As a result, judges will set a lax standard if you show them predominantly easy items and too harsh a standard if you show them mostly hard items. Too lax and too harsh are defined here in terms of the judge’s own intended cut score using a different set of items. Although the task group is correct that there is not as much research on the Bookmark procedure, it has the advantage that judges are shown items in the order of their scaled difficulties, so this essentially provides a scaffold to help judges avoid internal inconsistencies.³

Including more mathematicians and having judges take the test prior to setting standards are reasonable recommendations, but they will not necessarily improve the validity of resulting performance standards. The more significant of the task group’s recommendations is that international performance data should be brought to bear in the standard-setting process. Other normative and external validity evidence is needed as well. The National Research Council (Pellegrino et al., 1999) report also recommended that greater effort be made to acknowledge the judgmental nature of performance standards and to focus interpretation of assessment results more on change than on any absolute meaning of performance levels. Although it is fine to call for more research, my own belief is that we need better models of expertise rather than better procedures to shore up expert judgments. In most performance domains, expertise is a combination of proficient and flexible mastery of core knowledge and skills, plus some amount of specialized advanced knowledge. Experts do not all know all of the advanced knowledge. Therefore as tests move beyond basic skills, it becomes less and less satisfying to represent adept and advanced performance merely as a percentage correct on the total test. The current interest in learning progressions would be an example of a different way to approach the conceptualization of performance standards. Certainly we should be aware that averaging judges’ judgments is not likely to resolve the issue regardless of how sophisticated our processes become.

Multiple-Choice Versus Constructed Response Items

The Panel surprisingly reached the erroneous conclusion that multiple-choice and constructed response items measure the same mathematical competencies. Although the Task Group on Assessment acknowledged the prevalent study design that produced this result, the main report obscures this telling caveat. As shown in Table 1, the result of no difference comes almost entirely from studies where the researchers constrained the multiple-choice and open-ended items to be identical. Measurement researchers used the term *stem-equivalent* to indicate that the exact same question was asked in each pair of items, so the only difference was the provision of answer choices in the multiple-choice version. These

Table 1
Studies Comparing Multiple-Choice (MC) and Constructed Response (CR) Items Categorized by Study Design and Study Results

| Item Type | Results: Items Measure the Same | Results: Items Measure Differently |
|-----------------------|--|--|
| Stem-equivalent items | Behuniak, Rogers, & Dirir (1996): CR more difficult, fit with both 1- and 2-factor model. Gallagher (1992): No difference in strategy use between MC and CR. Hombo, Pashley, & Jenkins (2001): MC and grid-in items differed only in omit rates. Katz, Bennett, & Berger (2000): Some nontraditional solution strategies used with both MC and CR. Traub & Fisher (1977): MC and CR tap the same factor. | Birenbaum & Tatsuoka (1987): Open-ended items provide better diagnosis of misconceptions. Birenbaum, Tatsuoka, & Gutvitz (1992): Open-ended format provides more valid measure for diagnostic assessment. |
| Nonparallel items | Dossey, Mullis, & Jones (1993): CR items were more difficult and provided more information about proficient students. | Burton (1996): Women have a slight advantage on grid-in questions. DeMars (1998): Female students scored higher or relatively higher on CR scale. DeMars (2000): Students perform better on high-stakes than low-stakes test, but difference was larger for CR items. Garner & Engelhard (1999): Except for algebra items, MC items favored men and CR items favored women. Hastedt & Sibberns (2005): There is a clear tendency for CR to favor girls. Koretz, Lewis, Skewes-Cox, & Burstein (1993): The study only compared omit rates, which are higher for CR items. O'Neil & Brown (1998): Open-ended questions induced more cognitive strategy, less self-checking, and greater worry than MC questions did. Pollock & Rock (1997): Factor analysis showed differential difficulty of MC items compared to CR items for Black and Hispanic minority groups. |

researchers were interested in whether format per se made a difference in measuring students' abilities. Thus they carefully controlled for everything else, including content, cognitive process, and construct. The finding is essentially a tautology. Yes, if you strictly constrain multiple-choice and constructed response items to be identical, predictably they measure the same thing. The task group apparently admired the degree of control represented in such studies but failed to consider the measurement properties of constructed response items when these items are designed to address aspects of the content domain untapped by multiple-choice items.

In his 1984 article "The Real Test Bias," Norm Frederiksen debunked this body of literature and the conclusion that multiple-choice and constructed response items always measure the same thing. Leaving off the answer choices from multiple-choice questions, he noted, would not dramatically improve the cognitive level of multiple-choice items. He proposed, instead, a program of research where constructed response items designed to measure complex cognitive skills would then be turned into multiple-choice items. He conducted several studies using a formulating hypotheses format and found very weak correlations between the two "parallel"

forms of the tests. He also found striking differences between each of these tests in their relationship with other variables. For example, none of the scores from the multiple-choice form correlated with measures of ideational fluency, whereas the free-response scores, even with lower reliability, correlated substantially with number of ideas, number of unusual ideas, and number of high-quality unusual ideas.

To evaluate the task group's finding of no difference, we located copies of all but 1 of the 19 studies cited in the relevant section of the Task Group on Assessment chapter or appendix.⁴ Pollock, Rock, and Jenkins (1992) was not available electronically, and 2 of the studies cited did not directly compare multiple-choice and constructed response items (Bennett, Ward, Rock, & Lahart, 1990; Webb, 2001). The item design portion of each study was summarized along with results for item difficulty; correlations and factor analysis; and findings of group interactions by gender, ethnicity, country, or high-stakes context. Typically, constructed response items are more difficult, although it is possible to make multiple-choice items more difficult by choosing distractors to represent common misconceptions. Factor analyses were conducted in only

4 of the studies.⁵ Most famously, Traub and Fisher (1977) found only weak evidence of a format factor and concluded that in mathematics the two formats were equivalent. The other 3 studies, however, found evidence of fit for both one- and two-factor solutions (Behuniak, Rogers, & Dirir, 1996), distinct factor structures in the two tests analyzed separately (Birenbaum & Tatsuoka, 1987), and clear evidence of two distinct factors that also resulted in differential group performance (Pollock & Rock, 1997). Not surprisingly, the study that found separate factors for the two formats was the only one that did not constrain the items to be exactly parallel using the stem-equivalent study design.

In Table 1, a cross-tabulation shows the relationship between the item-design feature of each study and the effects on group performance. The stem-equivalent studies in the top row are the controlled studies that the Panel relied on to reach its conclusion that multiple-choice and constructed response items measure the same thing. Even here, two of the studies show important differences between the cognitive demands of the two types of tests. The bottom row shows the studies that used data from operational assessment programs, where constructed response items were purposely designed to tap aspects of the domain not tapped by multiple-choice items. Most of these studies are cited in the task group appendix, not in the chapter, and the focus of discussion is on group effects of the two different measures, not on what the tests measure. Yet, recurring group-by-format interactions can be compelling evidence that the two types of items are measuring different constructs. When such patterns are observed, there are essentially three possible explanations: differences in reliability, test bias, or true differences in the constructs being measured. The task group did not engage these issues, but often the authors of the original studies did. Although tests made of constructed response items are sometimes less reliable and the writing demand of some open-ended items could be considered a source of bias, the relative advantage of girls on constructed response items occurs frequently enough, in international studies for example, and across different open-ended formats that it is unlikely that the pattern can be attributed entirely to unreliability or construct-irrelevant variance.

Thus differences in conclusions about group differences—closing achievement gaps, for example, for girls and minority groups—should be taken seriously, as evidence that the substantive differences between the tests are important. The potential significance of the very real differences between multiple-choice and constructed response tests is perhaps best illustrated by Schmidt, Jakwerth, and McKnight's (1998) analysis of data from the Trends in International Mathematics and Science Study. In mathematics, country ranks changed appreciably when separate tests were constituted of only multiple-choice, only short-answer, or only extended response questions. Changes in country ranks ranged from 0 to 20, with an average change in rank of 5 places. Given the greater vulnerability of multiple-choice tests to teaching-the-test practice effects, it is especially worrisome that the Panel recommends that lower cost multiple-choice tests be treated as interchangeable with tests that include constructed response items.

Formative Assessment

In the chapter on instructional practices, the Panel makes several important points about the use of formative assessment. First and

foremost, the Panel emphasizes that ongoing monitoring of student learning is a hallmark of effective instruction. Based on its review of research, the Panel concludes that regular use of formative assessment improves student learning, especially if teachers receive guidance on how to use assessment results to individualize instruction. It calls for more research on the consequential validity of formative assessment tools (i.e., Do they really make teaching more effective?) and for research on the content and criterion-related validity of various classroom assessment tools. All of this is quite reasonable, yet a great deal is left out of this summary. The Panel explains that only one type of formative assessment has been studied with rigorous experimentation and that these assessments take between 2 and 8 minutes to administer. All of the discussion is about formal tools rather than instructional activities that might yield formative insights. Given members' expertise in mathematics and mathematics learning, it is frustrating that the Panel did not investigate the mathematical content of the formative assessments used in the studies it selected.

In the more extensive report of the Task Group on Instructional Practices, we are told that there are two distinct traditions of scholarship framing the study of formative assessment. One tradition represented by two National Research Council reports, *Adding It Up: Helping Children Learn Mathematics* (2001) and *How Students Learn: Mathematics in the Classroom* (2005), the task group characterizes as more informal and interactive but lacking any methodologically acceptable studies to examine its impact on student performance. Therefore, the task group relies exclusively on the second tradition, originating in school psychology and special education. The task group does not cite the famous review of formative assessment studies by Black and William (1998) nor more recent work relating the theory of formative assessment to research on learning (Shepard, 2006). This is unfortunate because it would have given the task group access to highly relevant experimental studies examining the effects of prior knowledge and feedback, as well as a more limited number of studies of self-assessment. For example, Kluger and DeNisi (1996) conducted a meta-analysis of 131 studies of feedback yielding 607 effect sizes and were able to explain what features of feedback are related to positive student outcomes. The research literature on transfer is also relevant, especially when one understands that the greatest worry about formative assessment in instructional interventions is that instructional materials are made to resemble outcome measures quite closely.

The Panel says that sociocultural learning theory is as yet unproven, but it accepts, without question, behaviorist assumptions about the sufficiency of test performance as proof of learning. Given the Panel's recognition that school districts have adopted various benchmark and interim assessments as formative tools specifically to improve performance on end-of-year tests, it is essential that the Panel's call for consequential validity studies include an alert to the threat of teaching-the-test effects. Studies designed to investigate whether formative assessment makes teaching more effective should be rigorously designed to ensure the adequacy of causal inferences, but this also means controlling for instrumentation and interactions of the treatment with instrumentation (Campbell & Stanley, 1963).

Summary

The Panel is at its best, even in the assessment chapter, when talking about mathematics content, and its main recommendations focus on the content of NAEP and state tests. Greater care should be taken to represent the Panel's Critical Foundations of Algebra—whole numbers, fractions, and particular aspects of geometry and measurement—in the items as well as the reporting strands of major tests. In addition, based on its expert review of items and the Daro et al. (2007) review of items by mathematicians, the Panel recommends that a more appropriate balance be sought in how algebra is “defined and assessed” (NMAP, 2008, p. xxv). Specifically, the Panel urges that assessment of algebra not include so many pattern problems. Like the Daro et al. study, which included analyses of state tests as well as NAEP, the Panel documented examples of “flawed” items where some aspect of wording, visual display, or context created sources of item difficulty unrelated to the intended mathematical content. Greater mathematics expertise is needed at both the item-writing and -review stages of test development to prevent these problems.

The Panel picked two important measurement issues to consider: the setting of performance standards and the psychometric properties of multiple-choice versus constructed response items. It is odd and disappointing that it did not consider other important issues such as the difficulties of creating coherent, focused assessments in the absence of a shared curriculum, and the problems of test score inflation and nongeneralized learning gains caused by teaching the test.

For any given research literature, there can be considerable variation in the methodological approaches taken to address particular questions. Some literatures proceed linearly from exploratory studies to more confirmatory, large-scale randomized trials, but this is not always the case. In some research literatures, focus on experimental controls leads to answering very different questions from those of most interest to the world of policy and practice. The experimental studies that compare multiple-choice and constructed response items constrain the two types of tests to be exactly parallel by taking away the answer choices as the only difference between the two types of items. By focusing on the no-difference findings from these studies, the Panel does a disservice to the field. In contrast, when constructed response items are purposely designed to measure something different, typically aspects that cannot be captured with multiple-choice tests, important differences are found. Thus the Panel has not really answered the critically important question as to whether higher cost constructed response items are needed to ensure valid representation of the content domain.

The Panel did not explain why it relied heavily on one simulation study (Reckase, 2006) as its source for making recommendations about standard-setting procedures. Reckase makes clear that his approach only addresses reliability issues, not validity. Because the Panel relied on this single study, its conclusions are quite misleading, which is unfortunate given the availability of more comprehensive treatment of standard-setting issues by a National Academy of Sciences Panel (Pellegrino et al., 1999). On the topic of formative assessment, the Panel does explain that controlled studies are the reason that it prefers studies coming

from special education and school psychology literatures rather than those from cognitive science. Unfortunately, not appreciating the theoretical arguments coming from the latter tradition, the Panel is unaware that relevant search terms such as *prior knowledge* and *feedback* would have turned up a host of relevant experimental studies. By focusing on a limited set of studies based on a particular version of formative assessment, the Panel implies that formative assessment requires formal tools (tests) with known psychometric properties that are administered and scored frequently. This is a very limited view of formative assessment, ignoring the possibility that well-designed instructional tasks can be used for formative assessment. At least in its recommendations for future research, the Panel calls for more rigorous investigation of more clinical types of formative assessment.

In the area of assessment, the Panel report provides a wealth of useful information, but readers will require a well-indexed traveler's guide to know what knowledge they can get accurately distilled from the report and what they will have to pursue elsewhere.

NOTES

¹Readers are reminded that on each major topic, the NMAP (2008) report must be read in five layers: Executive Summary, Chapter, Task Group Executive Summary, Task Group Report, and Appendix.

²There was no overlapping membership between the National Academy of Education and National Research Council committees, although the later National Research Council committee had access to the Academy studies.

³A refinement of the Bookmark procedure, called the Modified Mapmark method, was used recently to determine cut scores for the 2005 National Assessment of Educational Progress 12th Grade Mathematics Assessment. The Mapmark method goes even further in structuring the judgment task and provides feedback to judges based on item content domains to ensure that judgments stay consistent with the underlying psychometric scale.

⁴My thanks to Ph.D. student Kristen Davidson for her assistance with this analysis.

⁵A fifth study by O'Neil and Brown (1998) used confirmatory factor analysis to evaluate the structure of metacognitive and affective variables separately for students taking multiple-choice and open-ended items but did not directly compare structures for the two item types.

REFERENCES

- Behuniak, P., Rogers, J. B., & Dirir, M. A. (1996). Item function characteristics and dimensionality for alternative response formats in mathematics. *Applied Measurement in Education, 9*(3), 257–275.
- Bennett, R. E., Ward, W. C., Rock, D. A., & Lahart, C. (1990). *Toward a framework for constructed-response items* (ETS Research Rep. No. 90-7). Princeton, NJ: Educational Testing Service.
- Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385–395.
- Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16*(4), 353–363.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.
- Burton, N. W. (1996). Have changes in the SAT affected women's math scores? *Educational Measurement: Issues and Practice, 15*(4), 5–9.

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8*. Washington, DC: National Center for Education Statistics and American Institutes for Research.
- DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279–299.
- DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77.
- Dossey, J. A., Mullis, I. V. S., & Jones, C. O. (1993). *Can students do mathematical problem solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment* (No. 23-FR01). Princeton, NJ: Educational Testing Service.
- Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.
- Gallagher, A. (1992). *Strategy use on multiple-choice and free-response items: An examination of differences among high scoring examinees on the SAT-M* (No. ETS Research Rep. No. 92–54). Princeton, NJ: Educational Testing Service.
- Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29–51.
- Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states* (MG589). Santa Monica: RAND. Available at http://www.rand.org/pubs/monographs/2007/RAND_MG589.pdf
- Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*(2–3), 145–161.
- Herman, J. L. (2008). Accountability and assessment: Is public interest in K–12 education being served? In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 211–231). New York: Routledge Taylor & Francis.
- Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*(1), 3–17.
- Hombo, C. M., Pashley, K., & Jenkins, F. (2001). *Are grid-in response format items usable in secondary classrooms?* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.
- Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39–57.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.
- Koretz, D. (2008). Further steps toward the development of an accountability-oriented science of measurement. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 71–91). New York: Routledge Taylor & Francis.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.
- Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles: Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles.
- Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved July 21, 2008, from <http://epaa.asu.edu/epaa/v11n31/>
- National Center for Education Statistics. (2007). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES No. 2007–482). Washington, DC: Author.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.
- National Research Council. (2005). *How students learn: Mathematics in the classroom*. Washington, DC: National Academy Press.
- Olson, L. (2002, February 20). A “proficient” score depends on geography: Achievement levels vary widely by state. *Education Week, 21*(23), 1, 14.
- Olson, L. (2005, September 2). Defying predictions, state trends prove mixed on schools making NCLB targets. *Education Week, 25*(2), 1, 26–27.
- O’Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*(4), 331–351.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the Nation’s Report Card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.
- Pollock, J. M., & Rock, D. A. (1997). *Constructed response tests in the NELS: 88 high school effectiveness study* (No. NCES 97–804). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.
- Pollock, J. M., Rock, D. A., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response item formats in large-scale surveys*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standards setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4–18.
- Schmidt, W. H., Jakwerth, P. M., & McKnight, C. C. (1998). Curriculum sensitive assessment: Content *does* make a difference. *International Journal of Educational Research, 29*(6), 503–527.
- Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. East Lansing: U.S. National Research Center for the Third International Mathematics and Science Study, Michigan State University.
- Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement*. (4th ed., pp. 623–646). Washington, DC: National Council on Measurement in Education and American Council on Education/Praeger.
- Shepard, L. A. (2008). A brief history of accountability testing, 1965–2007. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 25–46). New York: Routledge Taylor & Francis.
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels*. Stanford, CA: National Academy of Education.

- Stecher, B., & Chun, T. (2001, November). *School and classroom practices during two years of educational reform in Washington state*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
- Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*(3), 355–369.
- Webb, D. C. (2001, August). *Classroom assessment strategies to help all students master rigorous mathematics*. Paper presented at a four-day workshop for District of Columbia Public Schools mathematics teachers, American Association for the Advancement of Science, Washington, DC.

AUTHOR

LORRIE A. SHEPARD is dean and professor in the School of Education at the University of Colorado, Boulder, Campus Box 249, Boulder, CO 80309; lorrie.shepard@colorado.edu. Her research focuses on psychometric topics such as validity theory, standard setting, and test bias, and evaluations of test use in contexts such as the identification of learning disabilities, kindergarten readiness screening, grade retention, teacher testing, high-stakes accountability testing, and formative assessment.

Manuscript received July 21, 2008
Revision received October 14, 2008
Accepted October 17, 2008