JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 47, NO. 1, PP. 6–24 (2010)

Test-Based Accountability: Potential Benefits and Pitfalls of Science Assessment With Student Diversity

Randall D. Penfield, Okhee Lee

School of Education, University of Miami, Coral Gables, Florida 33124-2040

Received 8 September 2008; Accepted 25 March 2009

Abstract: Recent test-based accountability policy in the U.S. has involved annually assessing all students in core subjects and holding schools accountable for adequate progress of all students by implementing sanctions when adequate progress is not met. Despite its potential benefits, basing educational policy on assessments developed for a student population of White, middle- and upper-class, and native speakers of English opens the door for numerous pitfalls when the assessments are applied to minority populations including students of color, low SES, and learning English as a new language. There exists a paradox; while minority students are a primary intended beneficiary of the test-based accountability policy, the assessments used in the policy have been shown to have many shortcomings when applied to these students. This article weighs the benefits and pitfalls that test-based accountability brings for minority students. Resolutions to the pitfalls are discussed, and areas for future research are recommended. © 2009 Wiley Periodicals, Inc. J Res Sci Teach 47: 6–24, 2010

Keywords: science; accountability, assessment, validity, reliability, bias, minority

Equal access to education for all students has been a cornerstone of educational policy in the U.S. since the passing of the Elementary and Secondary Education Act (ESEA) of 1965. Current educational policy, codified by the No Child Left Behind (NCLB, 2001) Act of 2001, is a test-based accountability policy where all students are held to the same rigorous academic standards in core subjects. The theory of action of the NCLB policy is based on mandated reporting of assessment scores for all students, such that schools, districts, and states are motivated to allocate resources in a way that gives all groups of students the chance of meeting the established standards.

Despite the anticipated beneficial properties of NCLB, numerous issues arise with respect to its effective implementation. Issues of unrealistic goals, curriculum contraction, test score validity, and inconsistent proficiency standards are just a few of many concerns voiced over NCLB (Kieffer, Lesaux, & Snow, 2008; Koretz, 2008; Kornhaber, 2008; Linn, 2008; Sunderman, Kim, & Orfield, 2005). While these concerns are pervasive across all students, issues of validity and fair use of test results are of heightened concern for minority students who have traditionally been underserved in K-12 educational settings. The concerns of test-based accountability systems with minority students are particularly relevant to science because science assessments often contain a high level of linguistically and culturally dependent content that can exacerbate the persistent gaps in science achievement and professions (National Center for Education Statistics, 2006).

In this article, we use the terms ''majority'' and ''minority'' students with reference to students' racial/ethnic, socioeconomic, and linguistic backgrounds. ''Majority'' is understood to refer not in a numerical sense, but rather in terms of social prestige, institutionalized privilege, and normative power. In classroom settings, ''majority'' students (i.e., those who are White, middle- or high-SES, and native speakers of English) are more likely than ''minority'' students (i.e., students of color, low-SES, and learning English as a new language) to encounter ways of talking, thinking, and interacting that are continuous with the skills and expectations they bring from home, and this continuity between home and school constitutes an academic

Correspondence to: R.D. Penfield; E-mail: penfield@miami.edu
DOI 10.1002/tea.20307
Published online 5 August 2009 in Wiley InterScience (www.interscience.wiley.com).

advantage relative to non-mainstream students. The more inclusive terms ''diverse student groups'' and ''students from diverse backgrounds'' are used to refer to the entire gamut of students, majority and minority. In addition, we denote particular student groups (e.g., economically disadvantaged, racial, and ethnic groups) using terms that are consistent with the language of NCLB legislation.

The measurement obstacles encountered in assessing science with minority students (i.e., issues of reliability, validity, and fairness) raise concern that test-based accountability policies may fall short in realizing their intended benefits for minority students. Although there is a small, but growing, body of research concerning science assessments with minority students (Luykx et al., 2007; Shaw, 1997; Siegel, 2007; Solano-Flores & Nelson-Barber, 2001), there has been little discussion of a broader range of issues concerning the reliability, validity, and fairness of measurement in science learning for minority students in the context of NCLB, how these measurement issues impede the realization of the intended benefits of NCLB for minority students, and directions for future research to improve the assessment of science for minority students in test-based accountability programs. This article aims to address this gap in the literature by: (a) weighing the potential benefits against the pitfalls of test-based accountability systems in science education with minority students, and (b) providing guidance for future research concerning the inclusion of science in test-based accountability systems in addressing student diversity.

In this article, we focus our attention on pitfalls associated with science assessments with minority students in the context of test-based accountability policies. We acknowledge that a variety of other potential pitfalls exist concerning the impact that test-based accountability policies have on the science curriculum, instruction, and student learning. In particular, NCLB is predicated on assumptions that the test-based policy yields improved science curricula and instruction (Aronson & Miller, 2007; Desimone, Smith, & Phillips, 2007; Geier et al., 2008; Lee & Luykx, 2005), and that the improved instruction positively impacts gains in science learning simply beyond higher test scores (Brickhouse, 2006; Champagne, 2006). Indeed, the extent to which such assumptions are met plays a pivotal role in determining gains in science learning. While we do not intend to detract from the importance of these assumptions underlying NCLB, our goal for this article is targeted to measurement issues specific to science assessment with minority students. We have restricted our attention to measurement issues for the following reasons: (a) the literature addressing assumptions concerning the impact of NCLB on science curriculum, instruction, and student learning is sparse at present due to the recency of implementing science assessment systems under NCLB, (b) unlike assumptions about curriculum, instruction, and student learning for which science educators have a wide range of and often extreme opinions, the science education research community generally does not address measurement issues, and (c) space limitations prohibit addressing assumptions across the areas of science curriculum, instruction, student learning, and assessment underlying NCLB in sufficient depth. It is our hope that presenting a comprehensive account of measurement issues particular to minority students, and weighing these issues against the potential benefits of NCLB, will stimulate discussion and debate concerning the nature of test-based accountability in science with minority students.

To meet the aim of this article, we have organized the contents in four sections. The first section provides a brief overview of the history of test-based accountability systems in the U.S. during the past few decades. The second section describes the potential benefits of test-based accountability with minority students by addressing what is gained from the perspectives of the student, the school, and the field of science education. The third section describes boundaries and pitfalls of test-based accountability with minority students by addressing measurement obstacles concerning validity, reliability, and fairness. The final section weighs the benefits and pitfalls of test-based accountability in science with minority students, and offers directions for future research addressing the boundaries and pitfalls of measurement issues in test-based accountability policies.

A Brief History of Test-Based Accountability

In 1965 President Lyndon B. Johnson passed ESEA, which sought to overhaul the federal funding policies of K-12 education in the U.S. The core of ESEA resided in the first provision, Title I, a program of federal aid for the education of low-SES children. Initially, Title I funding was allocated ''through a formula that ensured at least some money went to almost every school district in the nation'' (Hess & Petrilli, 2006, p. 9). The consequence of the widespread distribution of Title I funds among nearly all schools was that

it resulted in a limited impact for the intended students (Kantor & Lowe, 2006). It became evident that alternative policy was required to hold schools accountable for their effectiveness of teaching low-SES students and thus diminishing inequity.

It was the 1988 reauthorization of ESEA, signed by President Reagan, which introduced a focus on test-based accountability such that the performance of schools and districts receiving Title I funds would be monitored via annual student testing programs. The emphasis on assessment in accountability ''grew more focused and intensive with President Clinton's Goals 2000 initiative and the associated 1994 ESEA reauthorization, which mandated that states develop uniform academic standards for all their students and aligned assessments to measure student progress'' (Robelen, 2005, p. 42).

The passing of NCLB, the most recent reauthorization of ESEA, ushered in a new era of test-based accountability. NCLB is similar to Clinton's Goals 2000 initiative, but has broader jurisdiction across all students, plays a more forceful federal role in implementing the test-based accountability system, and has benefited from bipartisan support (Hess & Petrilli, 2006). Under NCLB, districts and schools are accountable for making an adequate level of gain in achievement each year. In the lingo of NCLB, the adequate level of gain is referred to as annual yearly progress (AYP). The theory of action of NCLB assumes that states, districts, and schools will allocate resources to best facilitate the attainment of AYP, which is determined by achievement test scores in reading and mathematics, graduation rates (for secondary schools), and at least one other academic indicator as determined by each state (e.g., writing or science). Thus, decisions concerning resources and practices are determined largely by test scores on state assessments in respective subjects. To accomplish this theory of action, NCLB introduces a three-stage process of accountability: (a) states define what constitutes AYP, (b) states measure achievement to determine whether AYP is met, and (c) sanctions are imposed on districts and schools if AYP is not maintained.

Although NCLB is perhaps most often associated with its test-driven accountability system, there is a second property that has garnered great attention and applause from the educational community and the public alike. NCLB mandates that each state report AYP disaggregated for the following student populations: (a) students of particular racial/ethnic groups as determined by the state (e.g., African American/Black, Hispanic/Latino, Asian/Pacific Islander); (b) students with limited English proficiency (LEP); (c) students with disabilities; and (d) students who are economically disadvantaged. Mandating this disaggregated reporting of AYP results in three desirable outcomes: (a) each of the groups is publicly monitored to examine achievement and progress, (b) resources are allocated differentially to these groups so that they meet AYP, and (c) if AYP is not met for these groups in schools receiving Title I funding, these groups are provided with additional resources such as supplemental educational resources (e.g., tutoring) and the right to transfer to another public school. Schools, districts, and states cannot hide historically underperforming demographic groups, since NCLB forces the state to publicly monitor these groups and to be accountable for their performance. In the end, lack of AYP by these groups is the responsibility of the state, district, and school, rather than the federal government.

While NCLB mandates reporting of AYP for reading and mathematics, the same has not been true for science. With respect to science, NCLB has required that by the 2007–2008 school year each state must have in place science assessments to be administered and reported for formative purposes at least once during grades 3–5, grades 6–9, and grades 10–12 (NCLB, §1111). As of 2009, NCLB had not required science to be included in AYP calculation. This is not to say that science assessments could not be included in the calculation of AYP, as a state could choose to include science in its AYP reporting by designating science as an additional academic indicator.

NCLB in science education involves multiple-steps for implementation in each state's assessment system: first, develop the assessment; second, administer the assessment; third, report the assessment results; fourth, decide whether to include the assessment results as part of state accountability, and finally, decide whether to include science in AYP calculation. The extent to which each state has met NCLB requirements pertaining to science assessment is unclear. What is known is that, as of 2008, nearly all states administer science assessments (R. Blank of the Council of Chief State School Officers, personal communication, August 29, 2008). What is not known is how many states report science assessment results or include these results as part of state accountability. In addition, while no accessible documentation exists concerning which states include science assessments in AYP calculation, a web-based search of AYP

indicators conducted in the fall of 2008 did not uncover any state that included science assessments in the calculation of AYP.

Although NCLB did not initially mandate that science be included in the calculation of AYP, the mandate of compulsory science assessments foreshadowed the potential future inclusion of science in AYP calculation. Indeed, the American Competitiveness Initiative proposed the future inclusion of science in AYP calculation (Domestic Policy Council, Office of Science and Technology Policy, 2006). The ultimate role that science will play in the test-based accountability systems in the near future hinges on decisions to be made in the reauthorization of NCLB.

Potential Benefits

The mandated science assessments in NCLB, and the proposal of future inclusion of science in AYP calculation, evidenced the pivotal role that science assessment does, and will, play in current and future test-based accountability systems. There are several potential benefits stemming from the inclusion of science assessment in NCLB mandates. As discussed in this section, major benefits include: (a) a concerted focus on reducing science achievement gaps between majority and minority student subgroups, (b) all students count in evaluations of school and district science achievement levels, and (c) science counts in determining whether adequate achievement levels have been attained. These potential benefits stem from the theory of action for NCLB that this policy will force states, districts, and schools to strategically allocate resources to science education for minority students who have traditionally performed poorly in science. It needs to be seen whether these potential benefits of including science in test-based accountability policy become actualized.

A Focus on Reducing Achievement Gaps

NCLB represented a shift in federal policy concerning educational improvement, particularly for minority students. Previous educational policy placed little emphasis on making schools and districts accountable for the disparity in school achievement between minority and majority students. Title I funds were distributed to provide resources earmarked for minority students, but little emphasis was placed on standardized measures to ensure that the resources resulted in improved educational outcomes for these students. In contrast, NCLB aims to close achievement gaps by identifying schools that have failed to provide ''students an opportunity to achieve the knowledge and skills described in the challenging academic content standards adopted by the State'' (NCLB, §1111) and by allocating resources for minority students in an effort to reduce the gaps.

Although the passing of NCLB was generally viewed as a victory for conservatives and their neo-liberal allies, NCLB also has received widespread support by many in the civil rights community (Kantor & Lowe, 2006). Sunderman (2008) summarizes this sentiment:

> For many in the civil rights community, NCLB represented an opportunity to focus on how public education has failed minority students. Skeptical that decisions made by state and local educators would result in tangible benefits for minority students, many civil rights advocates favored a stronger role for the federal government. That federal power had been successfully used to enforce civil rights and expand access to education for minorities, women, and students with disabilities led many to believe that federal power could be used to change educational practices and student learning. (p. 3)

The focus on reducing achievement gaps has positive implications for science, an academic area fraught with gaps in school achievement and professions among diverse student groups. Test-based accountability using standardized measures provides a system for evaluating the science achievement of diverse student groups and allocating resources to improve science instruction for minority students. It needs to be seen whether the focus on holding all students to the same academic standards can serve as a sufficient impetus to narrow achievement gaps in science.

All Students Count

Implicit in NCLB's emphasis on reducing achievement gaps is the requirement that all students count in school and district reporting of AYP. Provided that the number of students in each demographic group is

sufficient (as determined by each state), each school and district is required to report the status of AYP for each group. This ensures that minority students cannot be hidden and that their educational needs are given due attention. The incentive structure of NCLB will cause states, districts, and schools to reallocate resources to help students of all groups make AYP (Sunderman, 2008). The theory of action for NCLB is that because greater resources lead to improved opportunity to learn, it follows that test-based accountability systems can motivate schools to improve the opportunity to learn science for minority students.

Although science has been an optional component of AYP calculation, NCLB mandates that each state includes science in its assessment system and that the results of the science assessment be reported as disaggregated by subgroups. Reporting poor results for a state science assessment, even if science is not part of state accountability or AYP calculation, is an undesirable outcome for any school, district, or state. It needs to be seen whether the reporting of the results can influence the allocation of resources for the improvement of science for minority students.

Science Counts

Accountability policies influence instructional practices both in subject areas that are tested and in those that are not tested. When science is not included in accountability measures, it may be taught only minimally in the elementary grades (Knapp & Plecki, 2001; Spillane, Diamond, Walker, Halverson, & Jita, 2001). Even when science is included in accountability measures, it is generally tested at a particular grade level (e.g., fifth, eighth, or tenth grade; R. Blank of the Council of Chief State School Officers, personal communication, August 29, 2008). This contrasts reading and mathematics which are tested at every grade level. As a result, science still receives less attention than core subjects of reading and mathematics. In particular, schools serving English language learners (ELLs) and low-SES students are pressed to ensure basic proficiency in standard English literacy and numeracy, often at the expense of other subjects such as science (Lee & Luykx, 2005).

Separate from its relative standing to reading and mathematics, the theory of action for NCLB is based on the assumption that NCLB's mandated inclusion of science in each state's assessment system provides motivation for schools, districts, and states to allocate resources for science instruction (Lee & Luykx, 2005). There are four potential ways that NCLB can enhance the allocation of resources for science instruction. First, NCLB encourages schools to provide professional development in science that is critical for elementary school teachers who generally have inadequate preparation in science disciplines and science teaching. Second, by mandating the inclusion of science in each state's assessment system, NCLB motivates schools to employ a quality science curriculum aligned with state science content standards (Porter, 2002). Third, NCLB motivates schools to provide needed science materials and supplies, especially for hands-on science that is particularly effective with ELLs and students with limited formal school science (Lee & Fradd, 1998; Rosebery, Warren, & Conant, 1992). Finally, the mandated reporting of science assessment results may motivate schools to provide adequate instructional time for science, which is a significant hurdle given that science is often ignored in elementary schools, especially in urban schools where minority students tend to be concentrated (Lee & Luykx, 2005; Spillane et al., 2001). Again, it needs to be seen whether enhanced allocation of resources in these four areas of science instruction will become materialized.

Boundaries and Pitfalls

Although science test scores are not typically used in making decisions about student graduation and grade retention, the mandated reporting of science assessment results for NCLB does generate high stakes for districts, schools, and teachers. Given the consequences associated with science outcomes, it is important that the obtained scores hold up to professional standards of adequate psychometric and measurement properties. Documents outlining professional standards for educational testing include: (a) the 1999 Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), referred to as the *Standards* hereafter; (b) the fourth edition of Educational Measurement (Brennan, 2006); and (c) the Code of Fair Testing Practices (Joint Committee on Testing Practices, 2004). The remainder of this article draws heavily from these sources.

While science assessments may meet professional standard adequacy for majority students, there exists substantial risk that the threshold of acceptability may not be met for minority students. The reason for this risk is that these students have cultural and linguistic experiences that are potentially inconsistent with the properties of the test (e.g., item content) and the assessment process (e.g., science inquiry tasks in English). As such, there exist numerous boundaries to the appropriate implementation of science assessment for accountability with minority students. The following sections discuss such boundaries and obstacles with respect to three measurement issues: (a) validity, (b) reliability, and (c) fairness.

Validity Issues

Assessments use a sample of observations (i.e., responses to items or tasks) to make inferences about individuals' knowledge, skills, and abilities. At some point in the assessment process, one must consider whether the resulting interpretations and inferences are supported by evidence and theory. The process of accumulating evidence to form a scientifically based argument for score interpretation and use is known as validation (AERA/APA/NCME, 1999; Kane, 2006). A related term, validity, concerns the extent to which the accumulated evidence supports the proposed interpretations and uses of the scores obtained from the assessment.

A commonly misguided understanding of validity is that validity is a property of the assessment, such that we can speak of assessments as being valid (Frisbie, 2005). With rare exception (e.g., see Lissitz & Samuelson, 2007, p. 442) the technical measurement community rejects this notion of validity in favor of one that ascribes validity to the interpretations of scores generated by the assessment (Kane, 2006, 2008). That is, validity does not reside in the assessment itself, but rather in the interpretations and uses of scores. Ascribing validity to interpretations of scores rather than the assessments themselves has important consequences for the appropriate use of science assessments for diverse student groups—it allows for the possibility that scores may be valid for one student group (or one purpose) but not for another student group (or another purpose).

The issue of validity is of great concern for science assessments for diverse student groups. Given the language and content complexity inherent in science assessments, can science assessments yield valid interpretations and inferences of science achievement for students of differing cultures and varying levels of English proficiency? In this section we explore this issue, providing an overview of current research concerning the validity of scores obtained from science assessments for students of diverse backgrounds.

Language Complexity. Science assessments used in state-level accountability systems are typically developed for native speakers of English. This, in combination with the linguistic demands of tasks used to measure science knowledge and abilities, yields science assessments that contain a high level of linguistic complexity. While such science assessments may yield valid scores for majority students, there is substantial doubt that such science assessments yield valid scores for minority students, particularly the ELL population. The concern of validity for ELLs was a prominent theme of the report on appropriate test use by the National Research Council (1999), which stated, ''if a student is not proficient in the language of the test, her performance is likely to be affected by construct-irrelevant variance—that is, her test score is likely to underestimate her knowledge of the subject matter being tested'' (p. 225).

The impact that language issues may have on the validity of score interpretations for linguistic minority students is also supported by the *Standards* (AERA/APA/NCME, 1999):

> Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. In addition, language differences are almost always associated with concomitant cultural differences that need to be taken into account when tests are used with individuals whose dominant language is different from that of the test. (p. 91)

To address the issue of linguistic diversity with respect to assessment development and practices, the *Standards* (AERA/APA/NCME, 1999) devotes an entire chapter (chapter 9) to the issue, which includes a set of nine criteria for both test developers and users to consider with respect to linguistic issues in testing. Within these criteria reside several issues related to the validity of scores obtained for linguistic minority students, including: (a) the design of testing practices to reduce the threats to the validity of test score inferences that

may result from linguistic differences; (b) if appropriate, administration of the assessment in the test taker's most proficient language; and (c) provision of the requisite information for appropriate test use and score interpretation, such as manuals, instructions for score interpretation, and applicability of the test for non-native speakers of the language of the test.

The *Standards* (AERA/APA/NCME, 1999) and the National Research Council's report on appropriate test use (National Research Council, 1999) serve as evidence of a relatively wide-spread acknowledgement by the technical measurement community that linguistic complexity of assessments has the potential to threaten the validity of the obtained scores for ELLs. This issue poses particular concern for science due to the contextual and linguistic complexity of science assessments. Several recent studies have shed light on the jeopardized validity of assessments for ELLs. Solano-Flores and Trumbull (2003) examined how students' linguistic, cultural, and socioeconomic backgrounds interacted with the properties of items on a mathematics test to impact the way in which students interpreted and formed responses to the items. The results indicated that the linguistic properties of the items differentially impacted item interpretation for students of different linguistic, cultural, and SES backgrounds. Luykx et al. (2007) conducted a textual analysis of students' responses to a science assessment and found that their responses were influenced by linguistic factors: (a) non-standard spellings of English words that reflected the phonology of the students' home language but were often unintelligible to adult readers unfamiliar with their home language, and (b) semantic confusion concerning science terms (e.g., confusing *states* of matter with geopolitical states).

Culturally Dependent Content. Culture has a profound effect on the way in which people think about events, construct meanings, develop knowledge, solve problems, and communicate ideas. It follows that one's culture impacts how one interacts with a science assessment, and thus how the assessment elicits responses to items that are valid for making inferences about achievement. Solano-Flores and Nelson-Barber (2001) have described this aspect of validity as *cultural validity*, which they define as ''the effectiveness with which science assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of science items and respond to them'' (p. 555). They identified five areas in which the notion of cultural validity contributed to improving science assessment: student epistemology, student language proficiency, cultural worldviews, cultural communication and socialization styles, and student life context and values. The impact of culture on assessment responses was also investigated by Luykx et al. (2007) through a textual analysis of students' responses to a science assessment. They found cultural influences on responses to science assessments reflecting the knowledge, beliefs, or implicit assumptions deriving from students' homes or communities. They concluded that science assessments are inherently cultural objects, whose content and organization rely on cultural knowledge that different groups of students may not share.

The results of Solano-Flores and Nelson-Barber (2001) and Luykx et al. (2007) indicate that in order to obtain an accurate understanding of the validity of the inferences based on the science assessment scores for students from diverse cultural backgrounds, one must consider how cultural factors impact their responses to the tasks of the assessment. The issue of cultural validity is consistent with a recent emphasis by the measurement community on understanding the cognitive processes and strategies underlying examinee responses to assessment items (DiBello, Roussos, & Stout, 2007; Mislevy, 2006).

Real-life Experiences in Home and Community. One approach proposed to promote equitable assessment is to make assessments relevant to the knowledge and experiences of diverse student groups in their home and community environments. This approach focuses on the *content* of science assessments—to make assessments relevant to the knowledge and experiences that diverse student groups acquire in their home and community environments. Proponents argue that authentic tasks drawn from students' real-life situations, rather than decontextualized textbook knowledge, may motivate the students and enhance their performance (García & Pearson, 1994; Lacelle-Peterson & Rivera, 1994). Skeptics, however, claim that open-ended tasks may favor students with many opportunities to participate in science-rich environments over those lacking such opportunities (Hamilton, 1998; Shavelson, Baxter, & Pine, 1992). From this perspective, science assessment tasks based on classroom content and experiences are fairer than those requiring students to draw upon knowledge acquired outside the classroom. However, this holds only if all students have equal access to quality science instruction within the school environment, which is often not the

case since minority students tend to be concentrated in urban schools with limited funding and resources to support quality science instruction (Hewson, Kahle, Scantlebury, & Davies, 2001; Kahle, Meece, & Scantlebury, 2000; Spillane et al., 2001).

Performance Assessment Versus Multiple-Choice Formats. Another approach proposed to promote valid and equitable science assessments is to determine more effective *formats* for assessing student achievement in science. Traditional multiple-choice tests have been criticized for failing to measure the types of science knowledge and abilities that students should be expected to learn. Instead, alternative or performance assessments are called for, including open-ended or essay items, laboratory-based practical tests, portfolios, and opportunities to design and conduct experiments or projects (Ruiz-Primo & Shavelson, 1996). Advocates of performance assessment hold that it provides students with flexible and multiple assessment settings consistent with their cultural preferences, and permits students to communicate ideas in multiple ways (Darling-Hammond, 1994; García & Pearson, 1994; Lacelle-Peterson & Rivera, 1994). However, performance assessment tends to rely heavily on students' ability to read and write, confounding literacy skills with content knowledge. This is particularly problematic for ELLs, as well as for speakers of non-standard varieties of English (Ruiz-Primo & Shavelson, 1996; Shaw, 1997). Furthermore, regardless of their pedagogical value, performance assessment may be too costly and time-consuming to implement on a large scale (Stecher & Klein, 1997). As a result, large-scale assessment tends to rely on multiple-choice tests.

Reliability and Measurement Error Issues

Any assessment, regardless of the test content or the examinee, is implicitly imperfect in that a given examinee can obtain a range of possible test scores on any given administration of the test. This notion is perhaps most aptly summarized by Traub and Rowley (1991):

> To be concrete, let us suppose that the person is an eighth grade student and the characteristic of interest is achievement in mathematics. Further, let us imagine that this student can be retested many times with the same instrument, and that the student's mathematics achievement is *not* affected by the process . . . We would not expect all the scores resulting from this repeated testing experiment to be the same, just as we would not expect the repeated measurements that could be of the length of a table, using a tape-measure that has a suitably fine scale, to be all the same. In both cases, any variation in the number—scores, measurements—is presumed to be due to errors of measurement. (p. 173)

Thus, on any given administration of a test, a given examinee can obtain any one of numerous possible test scores, some scores being more probable than others. That there exists a range of possible test score values for a particular individual is a reflection of measurement error; the greater the range of possible values, the greater the degree of measurement error. The presence of substantial measurement error is a particularly undesirable state of affairs for an assessment, as asserted by the *Standards* (AERA/APA/NCME, 1999):

> Measurement error reduces the usefulness of measures. It limits the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process. Therefore, it reduces the confidence that can be placed in any single measurement. (p. 27)

There are many possible causes of the measurement error, including ''inconsistent behavior on the part of examinees, mistakes in marking, differential interpretations of examinees' responses, changes in an examinee's mood, motivation and mental alertness, guessing, and even sheer luck in the choice of questions for the test'' (Traub & Rowley, 1991, p. 173). In the context of diverse student groups, potential causes of measurement error may center on cultural and linguistic issues associated with the test content and the assessment procedures. These forms of measurement error pose particularly troubling dangers for science assessments due to the high contextual and linguistic demands of science assessments.

Because cultural and linguistic causes of measurement error are specific to particular student groups, it is expected that there may be greater measurement error for minority students than for majority students for whom the assessment is primarily developed. In the context of NCLB, where reporting group-level gains is

integral to the accountability process, a high degree of score reliability for minority students is critical. Substantial research indicates that high reliability for minority students is not always the case, as discussed next. While cultural and linguistic issues are both relevant to the study of measurement error, to date research has examined only the impact that language background has on measurement error. As a consequence, we focus our attention on the linguistic issues implicated in measurement error.

Linguistic Issues. Several studies have shown that the level of measurement error present in test scores varies as a function of ELL designation. Solano-Flores (2006) and Solano-Flores and Li (2006) showed that for a mathematics test with native speakers of Haitian-Creole classified as ELL, the results of a generalizability theory analysis of test scores revealed that substantial portions of the measurement error were attributable to the interaction of student, item, and code (language or dialect). Solano-Flores and Li (2006) summarized the results:

> Each test item poses a unique set of linguistic challenges and each student has a unique set of linguistic strengths and weaknesses. This sensitivity to language appears to take place at the level of dialect. Also, students from different speech communities within the same broad linguistic group may differ considerably in the number of items needed to obtain dependable measures of their academic achievement. (p. 13)

Similarly, Abedi (2004) showed the reliability of math, language, and science scores obtained from 9th grade assessments administered in multiple states differed for LEP and English only subpopulations (as the terms were used by the author). Although the reliability was higher for the English only group across all three subjects, one of the greatest discrepancies was observed in science, for which the reliability was 0.805 for the English only group compared to 0.597 for the LEP group. As a point of comparison, the reliability was 0.898 for the mathematics scores for the English only group compared to 0.802 for the LEP group. These results underscored the issue of inflated measurement error in science assessment scores for linguistic minority students.

Fairness Issues

An important issue in considering the appropriateness of science assessment for minority students in test-based accountability systems is the extent to which the resulting science test scores are fair, in so far as minority students have the same opportunity to succeed on the science test as their majority counterparts. A lack of fairness of science test scores could adversely impact the intended benefits of NCLB because the scores would not appropriately reflect the intended outcomes for minority students. The issue of fairness in high-stakes testing has gained increased attention over the past two decades, in part due to the increased reliance on standardized testing for individual accountability (e.g., promotion, graduation) and school-, district-, and state-level accountability, and in part due to the array of legal challenges to high-stakes testing (National Research Council, 1999; Phillips & Camara, 2006). Testaments to the increased visibility and importance allotted to fairness issues are that, for the first time ever, the issue of fairness has been devoted an entire chapter in the *Standards* (AERA/APA/NCME, 1999) and the fourth edition of Educational Measurement (Brennan, 2006), both sources serving as guides of best practices and contemporary trends in test design and use. The current landscape of fairness in testing has been shaped by ethical, legal, and psychometric issues, yielding a vision of fairness that is multi-faceted in nature. Components of fairness relevant to this discussion of the K-12 science assessment include: (a) item bias, (b) score equity analysis, and (c) opportunity to learn.

Item Bias and Differential Item Functioning. The issue of bias in testing has a long history of concern in both the general public and the technical measurement community (Zwick, 2002). In the measurement community, bias arises ''when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups'' (AERA/APA/NCME, 1999, p. 74). In the context of science assessment, the presence of bias indicates that students who have the same level of science proficiency, but belong to different groups (e.g., English only vs. ELL), do not have the same expected score on a science test. Bias can be the result of numerous factors, including: (a) language that

may have different interpretations for different student groups; (b) differential exposure to particular content that is secondary to the primary trait being measured (i.e., culturally specific content or language); and (c) language or content that is ''sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population'' (Bond, Moss, & Carr, 1996, p. 121). Given that items of science assessments contain large amounts of language and contextual information, the issue of item bias is highly relevant to science assessments for minority students.

Currently, the most widely used and psychometrically defensible approach for evaluating item-level bias is the framework of differential item functioning (DIF; Camilli & Shepard, 1994; Penfield & Camilli, 2007). To describe the concept of DIF, consider a particular item on a science test. Suppose that we estimate each student's science proficiency using their score on the science test (we acknowledge that this test score itself may be biased, but this a circularity with no current solution). If we compare students having the same test score, but belonging to different groups (e.g., English only vs. ELL), these two groups of students (matched on science proficiency) should have the same overall success rate on the item. We could make such a comparison for students at each test score, thus comparing students having similar proficiency in science, but belonging to different groups. If, at each test score, the success rate on the item was similar for the two groups of students, then we have evidence that students having the same science proficiency, but belonging to different groups, are performing equally on the item. In this situation, the item does not contain DIF and presumably is free of bias. However, if there is a consistently higher success rate for one group of students (e.g., English only) when compared to another group (e.g., ELL) matched on proficiency level, then we conclude that DIF exists (i.e., the item functions differentially for the two groups), and that there may exist a biasing factor (such as a linguistic or cultural factor) in the item causing students of one group (ELL) to underperform on the item when compared to their counterparts (English only) having the same overall science proficiency.

Numerous studies have demonstrated the presence of DIF associated with cultural and linguistic properties of item content in assessments of mathematics and language arts (Bolt & Ysseldyke, 2006; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Ferne & Rupp, 2007; O'Neil & McPeek, 1993). The study of DIF in science assessments, however, has been given substantially less attention in the literature. Despite the potential for the presence of culturally and linguistically biasing factors to arise in science assessments (due to the high contextual and linguistic demands of science assessments), relatively little attention has been given to the investigation of DIF in science assessments. Furthermore, the limited DIF research on science assessments focused on the comparison of males and females (O'Neil & McPeek, 1993; Zenisky, Hambleton, & Robin, 2003-2004), which is tangential to our current focus on minority students.

Two DIF studies have shown large DIF effects in science assessments for minority students. Nelson-Barber, Huang, Trumbull, Johnson, and Sexton (2008) found large DIF effects in a science assessment that appeared to be attributable to cultural and linguistic properties of the items causing problems for American Indian students relative to their White counterparts. Penfield, Alvarez, and Lee (2009) found large DIF effects in performance-based items of a science assessment that indicated differential difficulty of items between Hispanic and Black students. They made use of the differential step functioning (DSF; Penfield, 2007) framework, which pinpoints the specific score levels of performance-based tasks manifesting a potentially biasing factor. Using the DSF framework, they showed that the successful advancement (or step) to higher score levels of some performance-based tasks was differentially difficult for Hispanic and Black students (some steps favored Black students and some steps favored Hispanic students). This differential difficulty could be traced back to culture-specific content associated with the tasks defining the score levels. The results of these studies provide evidence that the items of science assessments contain cultural and linguistic properties that can lead to a bias of test scores for minority students.

Score Equity Analysis and Equating Invariance. Let us next turn our attention to the concept of score equity analysis based upon equating invariance. Measuring AYP is dependent on evaluating the growth between successive years. Because each year makes use of a different assessment to measure the achievement for that year (e.g., there is a different 5th grade science test each year in Florida), and because the difficulty of the tests varies across the years, the meaning of the test scores from year to year must be linked so that appropriate comparisons can be made and appropriate measures of AYP can be obtained. That is, the scales of

the tests from successive years must be linked so that the scores can be meaningfully compared (i.e., a score of 75 has the same meaning with respect to the standards across all years).

The process of linking the scores of different tests across the years such that the scores from these different tests can be used interchangeably is commonly referred to as *equating* (Holland & Dorans, 2006; Kolen & Brennan, 1995). Central to the process of equating is the development of an equating function, which is used to transform the scale of the scores on one test to the scale of the scores on a second test. It is possible, however, for the equating functions to vary across subpopulations of examinees, such that the simultaneous equating across all examinees using a single equating function would lead to biased scores for one or more subpopulations of examinees. The presence of equating function variability across subpopulations (a property known as a lack of population invariance) is a rather undesirable property if one wishes to equate two tests to appropriately measure growth for all examinees. As an example, the presence of unequal equating functions for two groups (e.g., English only vs. ELL) on a science test would indicate that a particular change in actual science test scores (say a growth of five test score units) would reflect different growth in actual science proficiency for the two groups. Dorans (2004) contended that between-group differences in equating functions indicate a lack of score equity across subpopulations, which jeopardizes the fairness of the assessments and the resulting consequences of the score interpretations.

Several studies have shown that between-group differences in equating functions exist as a function of gender and race/ethnicity in relation to college entrance exams (Dorans, 2004; Kobrin & Melican, 2007; Liu & Holland, 2008). Yet, little research has been conducted in relation to science assessment. The only published study of between-group differences in equating functions for a science assessment is a recent study by Yi and Harris (2008) that examined equating invariance in a science achievement test administered to high-school students. This study focused on groups defined according to science proficiency (i.e., science courses taken, science GPA), and found that differences in the equating functions existed for students who had and had not taken a physics course (presumably because taking a physics course is correlated with achievement). It is likely the case that not having taken a physics course would also be related to being low-SES and/or being a member of a racial/ethnic minority group, which provides evidence that equating functions for minority student groups may differ from that obtained when all students are lumped together into a single sample.

The research on score equity analysis via equating invariance is in its infancy, and much more research is required to gain an understanding of how equating functions may differ for diverse student groups in science assessments. The lack of current research, however, should not be interpreted as a lack of importance. Given the high linguistic and contextual demands of science assessments, we view this to be a critical issue in understanding whether science assessments used for measuring AYP lead to equitable and fair outcomes for minority students.

Equal Opportunity to Learn. A component of fairness for any assessment of content mastery is that the student has had adequate opportunity to learn the material being tested and to be successful in meeting the relevant academic standards as measured by the test. Adequate opportunity to learn is critical because it reflects the extent to which ''a student's performance on a test reflects knowledge and skill based on appropriate instruction or is attributable to poor instruction or to such factors as language barriers or construct-irrelevant disabilities'' (National Research Council, 1999, p. 275). The issue of opportunity to learn as a requisite property for fair testing is relevant to reading, mathematics, and science alike. Across all three of these subjects, a host of factors are involved in determining opportunity to learn, including quantity and quality of instruction, the alignment of test content and the curriculum received in class, the knowledge and abilities of each student, the interest and motivation of each student, resources available for effective learning, the learning climate, and class size (Darling-Hammond, 1992–1993; Elmore & Fuhrman, 1995; Levin, 2007; Starratt, 2003). In addition, science assessments typically involve items that draw heavily on academic English and highly contextualized scenarios, which further jeopardizes the opportunity that students have to master the requisite content and, perhaps more difficult, demonstrate this mastery on a timed test. In the context of state science assessments, the importance of opportunity to learn has been highlighted by the Committee on Test Design for K-12 Science Achievement (National Research Council, 2006):

> The fairness of assessments and the validity of the results depend on both the extent to which students have had the opportunity to learn the skills and material that are assessed and the use of assessments that are unbiased and accessible to a wide range of students with different abilities and disabilities. If students do not have the opportunity to learn the material or to demonstrate their knowledge in the context of appropriately designed assessments, it is impossible to know whether the results shed light on aspects of the curriculum, instructional strategies, or students' efforts or abilities, or whether they simply indicate that students have not been given a chance to learn what is being assessed or that the assessments are somehow not tapping into what they know in appropriate ways . . . As states make decisions about how to assess students' science literacy, they will need to consider the needs of English language learners and students with special needs and the challenges of devising technically sound accommodations for them. They will also need to consider the extent to which students with disabilities and English language learners have had an opportunity to learn the material covered by an assessment. (pp. 7–8)

The importance of opportunity to learn for the fair use of high-stakes tests has been addressed by two influential sources: (a) legal criteria of fair test use, as established by judicial decisions pertaining to the fair and non-discriminatory use of testing; and (b) professional criteria of fair test use, as documented in the *Standards* (AERA/APA/NCME, 1999) and the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004).

Let us first consider how legal criteria of fair testing practices in K-12 education have advanced the importance of opportunity to learn. Inadequate opportunity to learn has been a cornerstone of litigation pertaining to high-stakes tests arguing a violation of the Due Process Clause of the Fourteenth Amendment of the United States Constitution (Phillips, 2001; Phillips & Camara, 2006). Legal challenges to high-stakes testing have called upon two aspects of opportunity to learn: (a) curricular validity, and (b) fundamental fairness. First, curricular validity describes the extent to which the curriculum supports an opportunity for the students to attain the requisite skills to meet the relevant academic standards, as measured by the test. Evidence of curricular validity has been obtained from surveying teachers and students on the skills taught by the official curriculum (Debra P. v. Turlington, 1983), and reviews of textbooks and curricular materials (Phillips, 1996). Second, fundamental fairness is a somewhat broader issue concerning the ultimate opportunity for success on the test, independent of the curricular validity status (i.e., ultimate opportunity for success may be thwarted despite the presence of curricular validity). Fundamental fairness asserts that ''assessments must adhere to professional requirements, be valid, reliable and fair, avoid arbitrary or capricious procedures, and provide all students with conditions fostering an equal chance for success'' (Phillips, 1996, p. 7). Fundamental fairness requires that the test not be laden with procedures, language, or contextual information that places one or more students at a relative disadvantage, thus compromising the opportunity for success on the test (PASE v. Hannon, 1980).

Let us next consider how professional criteria of fair testing practices in K-12 education have advanced the importance of opportunity to learn. The *Standards* (AERA/APA/NCME, 1999) addresses this issue in criterion 13.5, which states that ''when test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn'' (p. 146). Although this criterion is written with respect to tests used to make decisions concerning promotion and graduation, it is important to remember that the *Standards* predates NCLB. In light of the high-stakes nature of NCLB to the student, teacher, school, and district, we can perhaps take liberty to interpret this criterion with a more inclusive definition of high-stakes that includes student-, teacher-, school-, and district-level accountability. In addition, the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004) addresses the issue of opportunity to learn by stipulating the following two principles: (a) if between-group differences in test performance arise, then the test users should ''determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed'' (p. 5); and (b) test developers or test users should ''inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies [making] such information available to all test takers'' (p. 10).

Despite the importance attached to opportunity to learn in legal and professional criteria of fair testing practices, opportunity to learn has not, in and of itself, been given sufficient leverage to prove a test to be

fundamentally unfair. While courts have acknowledged that some students may experience inadequate opportunity to learn due to little or no instructional match to the test, the courts have found that such cases may not rise to the level of a constitutional infringement (Van Horn Helvey, 2002, p. 30). Furthermore, following the court ruling in the landmark case of Debra P. v. Turlington (1983), courts have held that it is not constitutionally unfair that different students have teachers of different quality. The statement below from the *Debra P.* court underscores the limited weight that opportunity to learn has been given in decisions concerning test fairness:

> But what does the Constitution require in this instance? It may not be fair to expect students with differing interests and abilities to learn the same material at the same rate, but is it unconstitutional? Similarly, it may be inequitable that some students, through random selection, are assigned to mediocre teachers while others are given excellent instructors, but does this rise to the level of a constitutional violation? . . . Suppose that there is one student who never encountered a teacher who taught the [appropriate] skills, or a teacher who taught the skills well, would the entire test be declared invalid? What if the number of students complaining was 3,000 rather than one? (pp. 183–184)

By its very nature, NCLB holds all students to the same standards, so that schools, districts, and states are motivated to provide all students with quality instruction. Yet, research indicates glaring disparities in opportunity to learn with respect to teacher quality (Darling-Hammond, 2004), curricular resources (Darling-Hammond, 2004), and English proficiency (Rumberger & Gandara, 2004). This raises a critical question: Is it appropriate and justifiable to implement a standards-based accountability program in the absence of adequate opportunity to learn for all students? There exists no clear-cut answer to this question, as there are arguments on both sides of the issue. On the one hand, making high-stakes decisions without verifying adequate opportunity to learn seems unfairly punitive; on the other hand, implementing standards for opportunity to learn adds on layers of undesirable regulations (Elmore & Fuhrman, 1995; Porter, 1995; Starratt, 2003). Short of any unifying answer to the question at hand, we can only offer the following conclusion—opportunity to learn science content and to display mastery of the content on an assessment can vary dramatically as a function of linguistic, cultural, and SES backgrounds, and thus the fair use of science assessments for test-based accountability systems must be sensitive to this effect.

Discussion

The NCLB Act of 2001 is unprecedented in terms of its broad jurisdiction of test-based accountability. It is also unprecedented in terms of its expectation of high academic standards for all students and reporting of AYP for minority students who have traditionally been underserved in the education system. While reading and mathematics have been the core subjects in education reform during the past several decades, science has been recently added as part of high-stakes state assessments and accountability systems, and has been considered for future inclusion in AYP. While the historically unprecedented attention to science as part of test-based accountability systems brings a heightened level of attention to the science education community, it also presents challenges in implementing equitable and fair assessments of minority students. Although the science education community can learn lessons from the reading and mathematics education communities, high contextual and linguistic demands of science assessments pose a unique set of difficulties. Thus, it is critically important that the science education community, the measurement community, the policy community, and the legal community work collaboratively in order to improve the educational equity resulting from test-based accountability policies.

Weighing the Benefits and Pitfalls

Test-based accountability in science with minority students has several potential benefits. One potential benefit is the focus on reducing achievement gaps between student subgroups. The persistent gaps in science achievement have been demonstrated since the inception of NAEP in 1969 (National Center for Education Statistics, 2006). Test-based accountability in science for all students has the potential to reduce these gaps. Another potential benefit is that all students count. Test-based accountability in science obliges schools, districts, and states to attend to the science achievement of students of historically low-performing

groups, and provides motivation for the education system to address their educational needs. Still another potential benefit is the integration of science in the school curriculum. Science has traditionally been ignored, especially in urban schools where minority students tend to be concentrated, due to the perceived urgency of developing basic literacy and numeracy (Lee & Luykx, 2005). Test-based accountability in science has the potential to motivate schools and districts to allocate resources and instructional time for science.

However, these potential benefits of test-based accountability in science with minority students can be accomplished only under certain conditions. First, science education must receive more attention in terms of quality science curriculum, professional development of teachers, materials and supplies, and instructional time for science. Second, states, districts, and schools must reallocate resources for traditionally underserved groups to learn rigorous science standards and make adequate academic growth across successive years until science achievement gaps are reduced and eventually closed. Third, all students must have the opportunity to learn science, and this is particularly important for minority students who have traditionally performed poorly in science. Finally, science assessments must hold up to professional standards of psychometric and measurement properties in terms of validity, reliability, and fairness (AERA/APA/NCME, 1999; Brennan, 2006; Joint Committee on Testing Practices, 2004). If these conditions are not met, the test-based accountability in science may exasperate inequalities already present in the educational system.

The discussion presented herein suggests a clear risk that the potential benefits of including science in NCLB for minority students will be attenuated by a lack of psychometric adequacy of the science assessments for minority students. In particular, the less than ideal psychometric properties of the assessments for minority students jeopardize the accuracy and appropriateness of the inferences of science knowledge and abilities made about these students. The presence of construct-irrelevant variance arising from linguistic and cultural properties of test content unveils the possibility of a systematic bias in the scores for minority students. This systematic bias not only confounds the interpretation of test scores for diverse student groups within a given year, but impacts the measure of AYP because the equating function used to link the scores between adjacent years may not work the same way for diverse student groups as it does for the entire student population as a whole. That is, a given magnitude of measured AYP may reflect different magnitudes in actual growth in proficiency between minority and majority student groups. Although minority students are often intended beneficiaries of test-based accountability policy by the allocation of resources, the same students are at the greatest risk of experiencing a systematic bias in the measure of AYP.

Directions for Future Research

Although our conclusion concerning current test-based accountability policies is that psychometric challenges associated with measuring science achievement of minority students may attenuate the potential benefits of NCLB for these students, we also see several lines of research that can aid in removing some of the current barriers to the intended benefits. The first line of research deserving greater attention concerns studies aimed at improving our understanding of why and how science assessments have different measurement properties for minority students than their majority counterparts. Given the preponderance of evidence indicating that the validity, reliability, and fairness of science assessments are jeopardized for diverse student groups due to cultural and linguistic properties of the assessment, we need a broader understanding of how construct-irrelevant variance exists in science assessments with diverse student groups, and how the presence of such construct-irrelevant variance generates systematic biases in assessment scores and the measure of AYP for particular student groups. This calls for studies of how and why individual test items function differently for different student groups (i.e., the study of differential item functioning and differential step functioning; Penfield & Camilli, 2007; Penfield, 2007) and studies of how between-group differences in test equating (i.e., equating invariance; Dorans, 2004) leads to systematic biases in the measure of AYP. We need a better understanding of what assessment components (e.g., tasks, item types, wording, and content) are differentially difficult for minority students. This research will help us better understand how science assessments are dependent on the linguistic and cultural background of the student.

Second, in addition to understanding why and how linguistic and cultural factors can adversely impact the validity and reliability of test scores of minority students, it is important to pursue research addressing how science assessments can be modified to be sensitive to linguistic and cultural factors. For example, accommodation strategies that reduce the linguistic complexity of science assessments may lessen the impact

of linguistic diversity on science assessment scores, a result that has been demonstrated in mathematics assessment (Abedi & Lord, 2001). Other accommodation strategies have also been recommended for ELLs, and these have shown varying levels of success in narrowing the achievement gap between ELLs and non-ELLs (Abedi, Hofstetter, & Lord, 2004). Future research may benefit from examination of accommodation strategies in science assessments with ELLs as well as students from culturally diverse backgrounds. This line of inquiry will shed light on how the use of accommodations for minority students can aid in improving the psychometric properties of science assessments, which may improve the validity of test scores used in the calculation of AYP.

A third line of research involves developing multiple forms of assessments to examine whether different forms have linguistic, cultural, and contextual information specific to the particular student population being tested. Each form could contain a core set of items common to all forms (or some variation of different overlapping items across multiple forms), as well as a set of items that have unique linguistic and cultural properties targeted to each particular student population (i.e., each minority group) being tested. The scores from the multiple forms of the assessment can be equated through the core set of common items. Given the high level of psychometric sophistication that the field of item response theory (IRT; Lord, 1980) has brought to test equating (Kolen & Brennan, 1995), equating across multiple forms may be a viable approach to obtaining scores on a common metric that are optimally valid for both majority and minority students. This idea is a variation of a recommendation proposed by Solano-Flores, Lara, Sexton, and Navarrete (2001) that students with limited English proficiency be assessed using test items in both their first language and English to inform test developers ''how students interpret items and what kind of thinking the items elicit in each language'' (pp. 22–23). This approach may improve the validity of test scores used in the calculation of AYP.

Finally, future research may address the issue of opportunity to learn in the context of science education with students of diverse backgrounds. The enactment of test-based accountability in science with diverse student groups necessitates careful examination of current practices in science curriculum, instruction, and teacher professional development with particular attention to establishing the extent to which current instructional practices yield opportunity to learn for minority students. Since science was traditionally not included in accountability measures (and still is not part of AYP), research and development efforts in science education have not received high priority, compared to other core subjects of reading, mathematics, and even writing. The negative impact of educational policies affecting science education tends to be greater for minority students who tend to be concentrated in urban schools. For example, if high-quality instructional materials that meet current science education standards are difficult to find (National Science Foundation, 1996), materials that also take into account the cultural and linguistic diversity of today's classrooms are even scarcer (National Science Foundation, 1998). It follows that resolutions to the issue of fairness are embedded not only in making science assessments that are more sensitive to linguistic and cultural diversity of the examinee population, but also in providing opportunity to learn for all students (Lee & Fradd, 1998; Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008). We contend that until opportunity to learn is ensured for minority students, there will be an unavoidable and undeniable violation of fundamental fairness of any test-based accountability system. This assertion underscores the importance of additional research related to the development of effective ways to provide the opportunity to learn science to minority students through quality science curriculum, professional development opportunities for teachers, and adequate instructional time for science.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. Educational Researcher, 33, 4–14.

Abedi, J., Hofstetter, C.H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. Review of Educational Research, 74, 1–28.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. Applied Measurement in Education, 14, 219–234.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: APA.

Aronson, I., & Miller, J. (2007). Competing horizons. The Science Teacher, 74(7), 64–67.

Bolt, S.E., & Ysseldyke, J.E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. Applied Measurement in Education, 19, 329–355.

Bond, L., Moss, P., & Carr, P. (1996). Fairness in large-scale performance assessment. In G.W. Phillips & A. Goldstein (Eds.), Technical issues in large-scale performance assessment (pp. 117–140). Washington, DC: National Center for Education Statistics.

Brennan R.L. (Ed.). (2006). Educational measurement (4th edn.) Westport, CT: American Council on Education and Praeger.

Brickhouse, N.W. (2006). Celebrating 90 years of science education: Reflections on the gold standard and ways of promoting good research. Science Education, 90, 1–7.

Camilli, G., & Shepard, L.A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.

Champagne, A. (2006). Then and now: Science assessment 1996–2006. School Science and Mathematics, 106, 113–123.

Darling-Hammond, L. (1992-1993). Creating standards of practice and delivery for learner-centered schools. Stanford Law and Policy Review, 4, 37–48.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. Harvard Educational Review, 64, 5–30.

Darling-Hammond, L. (2004). Inequality and the right to learn: Access to qualified teachers in California's public schools. Teacher's College Record, 106, 1936–1966.

Debra P. v. Turlington, 564 F. Supp. 177, 186 (M.D. Fla. 1983).

Desimone, L.M., Smith, T.M., & Phillips, K.J.R. (2007). Does policy influence mathematics and science teachers' participation in professional development? Teachers College Record, 109, 1086–1122.

DiBello, L.V., Roussos, L.A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In S. Sinharay & C.R. Rao (Eds.), Handbook of statistics, Volume 26: Psychometrics (pp. 979–1030). New York: Elsevier.

Domestic Policy Council, Office of Science and Technology Policy. (2006). American competitiveness initiative: Leading the world in innovation. Retrieved August 11, 2008, from www.nist.gov/director/reports/ACIBooklet.pdf.

Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity. Journal of Educational Measurement, 41, 43–68.

Elmore, R.F., & Fuhrman, S.H. (1995). Opportunity-to-learn standards and the state role in education. Teachers College Record, 96, 432–457.

Ercikan, K., Gierl, M.J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. Applied Measurement in Education, 17, 301–321.

Ferne, T., & Rupp, A.A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. Language Assessment Quarterly, 4, 113–148.

Frisbie, D.A. (2005). Measurement 101: Some fundamentals revisited. Educational Measurement: Issues and Practice, 24(3), 21–28.

García, G.E., & Pearson, P.D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), Review of research in education: Vol. 20 (pp. 337–391). Washington, DC: American Educational Research Association.

Geier, R., Blumenfield, P.C., Marx, R.W., Krajcik, J.S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. Journal of Research in Science Teaching, 45, 922–939.

Hamilton, L.S. (1998). Gender differences on high school science achievement tests: Do format and content matter? Educational Evaluation and Policy Analysis, 20(3), 179–195.

Hess, F.M., & Petrilli, M.J. (2006). No child left behind primer. New York: Peter Lang.

Hewson, P.W., Kahle, J.B., Scantlebury, K., & Davies, D. (2001). Equitable science education in urban middle schools: Do reform efforts make a difference? Journal of Research in Science Teaching, 38(10), 1130–1144.

Holland, P.W., & Dorans, N.J. (2006). Linking and equating. In R.L. Brennan (Ed.), Educational measurement. 4th edn. (pp. 187–220). Westport, CT: American Council on Education and Praeger.

Joint Committee on Testing Practices. (2004). Code of fair testing practices in education. Washington, DC: Author.

Kahle, J.B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? Journal of Research in Science Teaching, 37(9), 1019–1041.

Kane, M. (2006). Validation. In R.L. Brennan (Ed.), Educational measurement (4th edn., pp. 17–64). Westport: CT: American Council on Education and Praeger.

Kane, M. (2008). Terminology, emphasis, and utility in validation. Educational Researcher, 37, 76–82.

Kantor, H., & Lowe, R. (2006). From new deal to no deal: No Child Left Behind and the devolution of responsibility for equal opportunity. Harvard Educational Review, 76, 474–502.

Kieffer, M.J., Lesaux, N.K., & Snow, C.E. (2008). Promises and pitfalls: Implications of NCLB for identifying, assessing, and educating English language learners. In G.L. Sunderman (Ed.), Holding NCLB accountable: Achieving accountability, equity, and school reform (pp. 57–74). Thousand Oaks, CA: Corwin Press.

Knapp, M.S., & Plecki, M.L. (2001). Investing in the renewal of urban science teaching. Journal of Research in Science Teaching, 38, 1089–1100.

Kobrin, J.L., & Melican, G.J. (2007). Comparability of scores on the new and prior versions of the SAT Reasoning Test (College Board Research Note RN-31). New York: The College Board.

Kolen, M.J., & Brennan, R.L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Koretz, D. (2008). The pending reauthorization of NCLB: An opportunity to rethink basic strategy. In G.L. Sunderman (Ed.), Holding NCLB accountable: Achieving accountability, equity, and school reform (pp. 9–26). Thousand Oaks, CA: Corwin Press.

Kornhaber, M.L. (2008). Beyond standardization in school accountability. In G.L. Sunderman (Ed.), Holding NCLB accountable: Achieving accountability, equity, and school reform (pp. 43–55). Thousand Oaks, CA: Corwin Press.

Lacelle-Peterson, M.W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. Harvard Educational Review, 64, 55–75.

Lee, O., & Fradd, S.H. (1998). Science for all, including students from non-English language backgrounds. Educational Researcher, 27(3), 12–21.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up educational innovations with nonmainstream students in elementary school science. American Educational Research Journal, 43, 411–438.

Lee, O., Maerten-Rivera, J., Penfield, R.D., LeRoy, K., & Secada, W.G. (2008). Science achievement of English language learners in urban elementary schools: Results of a first-year professional development intervention. Journal of Research in Science Teaching, 45, 31–52.

Levin, H.M. (2007). On the relationship between poverty and curriculum. North Carolina Law Review, 85, 1381–1417.

Linn, R.L. (2008). Toward a more effective definition of adequate yearly progress. In G.L. Sunderman (Ed.), Holding NCLB accountable: Achieving accountability, equity, and school reform (pp. 27–42). Thousand Oaks, CA: Corwin Press.

Lissitz, R.W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36, 437–448.

Liu, M., & Holland, P.W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. Applied Psychological Measurement, 32, 27–44.

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

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. Teachers College Record, 109, 897–926.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), Educational measurement (4th edn., pp. 257–305). Westport, CT: American Council on Education and Praeger.

National Center for Education Statistics. (2006). The nation's report card: Science 2005. Washington, DC: Government Printing Office.

National Research Council. (1999). High stakes: Testing for tracking, promotion, and graduation. Washington, DC: National Academy Press.

National Research Council. (2006). Systems for state science assessment. Washington, DC: National Academies Press.

National Science Foundation. (1996). Review of instructional materials for middle school science. Washington, DC: Author.

National Science Foundation. (1998). Infusing equity in systemic reform: An implementation scheme. Washington, DC: Author.

Nelson-Barber, S., Huang, C.-W., Trumbull, E., Johnson, Z., & Sexton, U. (2008, March). Elicitory test design: A novel approach to understanding the relationship between test item features and student performance on large-scale assessments. Paper presented at the annual meeting of the American Educational Researcher Association, New York.

No Child Left Behind Act of 2001, Public Law No. 107-110.

O'Neil, K.A., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 255–276). Hillsdale, NJ: Erlbaum.

PASE v. Hannon, 506 F. Supp. 831 (N.D. Ill. 1980).

Penfield, R.D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. Journal of Educational Measurement, 44, 187–210.

Penfield, R.D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning form to improve the interpretation of DIF in polytomous items. Applied Measurement in Education, 22, 61–78.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C.R. Rao (Eds.), Handbook of statistics, Volume 26: Psychometrics (pp. 125–167). New York: Elsevier.

Phillips, S.E. (1996). Legal defensibility of standards: Issues and policy perspectives. Educational Measurement: Issues and Practice, 15(2), 5–13, 19.

Phillips, S.E. (2001). GI Forum v. Texas Education Agency: Psychometric evidence. Applied Measurement in Education, 13, 343–385.

Phillips, S.E., & Camara, W.J. (2006). Legal and ethical issues. In R.L. Brennan (Ed.), Educational measurement (4th edn., pp. 733–755). Westport, CT: American Council on Education and Praeger.

Porter, A.C. (1995). The uses and misuses of opportunity-to-learn standards. Educational Researcher, 24(1), 21–27.

Porter, A.C. (2002). Measuring the content of instruction: Uses in research and practice. Educational Researcher, 31(7), 3–14.

Robelen, E.W. (2005). 40 years after ESEA, federal role in schools is broader than ever. Education Week, 24(31), 1, 42.

Rosebery, A.S., Warren, B., & Conant, F.R. (1992). Appropriating scientific discourse: Findings from language minority classrooms. Journal of the Learning Sciences, 21, 61–94.

Ruiz-Primo, M.A., & Shavelson, R.J. (1996). Rhetoric and reality in science performance assessments: An update. Journal of Research in Science Teaching, 33, 1045–1063.

Rumberger, R., & Gandara, P. (2004). Seeking equity in the education of California's English learners. Teachers College Record, 106, 2032–2056.

Shaw, J.M. (1997). Threats to the validity of science performance assessments for English language learners. Journal of Research in Science Teaching, 34, 721–743.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 24(4), 22–27.

Siegel, M.A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. Journal of Research in Science Teaching, 44, 864–881.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. Teachers College Record, 108, 2354–2379.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). Testing English language learners: A sampler of student responses to science and mathematics items. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. Educational Measurement: Issues and Practice, 25(1), 13–22.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. Journal of Research in Science Teaching, 38, 553–573.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. Educational Researcher, 32(2), 3–13.

Spillane, J.P., Diamond, J.B., Walker, L.J., Halverson, R., & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. Journal of Research in Science Teaching, 38, 918–940.

Starratt, R.J. (2003). Opportunities to learn and the accountability agenda. Phi Delta Kappan, 85, 298–303.

Stecher, B.M., & Klein, S.P. (1997). The cost of performance assessments in large-scale testing programs. Educational Evaluation and Policy Analysis, 19, 1–4.

Sunderman, G.L. (2008). Introduction: Rethinking the challenge of NCLB. In G.L. Sunderman (Ed.), Holding NCLB accountable: Achieving accountability, equity, and school reform (pp. 1–8). Thousand Oaks, CA: Corwin Press.

Sunderman, G.L., Kim, J.S., & Orfield, G. (2005). NCLB meets school realities: Lessons from the field. Thousand Oaks, CA: Corwin Press.

Traub, R.E., & Rowley, G.L. (1991). An NCME instructional module on understanding reliability. Educational Measurement: Issues and Practice, 10(1), 171–179.

Van Horn Helvey, S. (2002). Equation protection/Title VI and due process in high-stakes testing: An examination of the Chicago Public Schools' promotion policy. Children's Legal Rights Journal, 22, 23–34.

Yi, Q., & Harris, D.J. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. Applied Psychological Measurement, 32, 62–80.

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2003–2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. Educational Assessment, 9, 61–78.

Zwick, R. (2002). Fair game? The use of standardized admissions tests in higher education. New York: RoutledgeFalmer.