Performance Assessment and Diversity

Jeffery P. Braden
University of Wisconsin–Madison

The use of performance assessments with ethnically and linguistically diverse students raises five questions: (a) How do minority groups score on performance assessments relative to majority groups? (b) Do performance assessments enhance instruction for diverse students? (c) Are common performance standards good for diverse students? (d) Can (and should) performance assessments be translated into other languages? and (e) Are performance assessments biased when used with diverse groups? This article reviews literature to address these points and identifies areas of consensus and controversy. Although performance assessments share many of the same difficulties as traditional assessments (e.g., score inequities among groups), they also promise richer assessment more closely aligned to educational standards than traditional achievement tests.

Performance assessment is a new set of assessment technologies based on an old set of ideas. Performance assessment (also known as authentic assessment) seeks to measure complex, student-constructed performances to determine what students do and do not know. The use of complex tasks to measure student performance was largely replaced by objective, selected response (i.e., multiple-choice) tests following World War I. Performance assessment reintroduces complex student performances (e.g., essays, scientific experiments, practical and creative solutions to problems) to assess educational outcomes (Cizek, 1991).

The primary impetus for the rebirth of performance assessment is the standards-based school reform movement. In the past 15 years, many state education agencies have adopted educational standards, with the strong support of the past two U.S. Presidents. These standards require students to draw on knowledge and skills across academic domains and levels of complexity and to apply their knowledge, skills, and abilities to solve complex, realistic problems. Many educators believe multiple-choice tests cannot measure complex outcomes, and so at least 40 states (Elliott, 1995) are developing and/or have adopted performance assessment technologies to assess student acquisition of educational standards.

Address correspondence to Jeffery P. Braden, Dept. of Educational Psychology, 1025 W. Johnson St., Madison, WI 53706-1796. Tel.: (608) 262-4586 (u/TDD). E-mail: jbraden@education.wisc.edu.
I will explore the intersection of diversity and performance assessment. As I define it, diversity includes ethnic, cultural, and linguistic differences, but not disability status and individual differences. Although performance assessment poses significant issues for diversity associated with individual differences and disability, these issues are discussed elsewhere (Coffman, 1993; McDonnell, McLaughlin, & Morrison, 1997). My discussion begins by defining performance assessment and concludes with how performance assessments apply to ethnically, culturally, and linguistically diverse groups.

DEFINING PERFORMANCE ASSESSMENT

Performance assessment uses complex tasks that require students to understand problems, accurately apply their knowledge and skills to develop a solution to the task, and communicate their problem-solving process and solution. Most often, tasks are developed to elicit evidence of student competencies related to one or more academic standards. For example, a science task might assess students’ ability to “recognize, define, and solve a science problem and communicate the strategies used in appropriate scientific fashion” (Wisconsin Department of Public Instruction, 1995, p. 27). Students respond to the task by writing a narrative, drawing a picture or graph, measuring and computing appropriate values, and the like. Experts (teachers) review student responses and rank them using a scoring guide (i.e., rubric). Rubrics classify performance in an ordinal scale to establish proficiency on the standard. Categories typically range from minimal to advanced proficiency, often with two to seven values or levels. Schools typically report the proportion or number of students in each proficiency category in aggregate to stakeholders. Some districts/states may also release individual student scores, whereas others do not. Performance assessments are often part of school-, district-, and state-wide accountability programs and may have high-stakes uses for schools (e.g., budget allocations) and individual test takers (e.g., promotion to the next grade, high school graduation). Ideally, teachers should develop and use informal performance assessments in their classrooms to guide instruction and ongoing student evaluation so that students are well-prepared to demonstrate mastery of academic standards on formal performance assessments.

Performance versus Traditional Assessment

Performance assessments differ from traditional academic achievement tests in many ways. Some of the dimensions along which performance and traditional assessments can be compared are listed in Table 1. In addition to the features suggested, other terms have been suggested to distinguish performance from traditional assessments. “Authentic” describes complex tasks that (a) have meaning and value beyond school (e.g., running a small business); (b) require constructed responses; and (c) require synthesis and integration of knowledge and
skills across domains. "Portfolios" are long-term collections of work, typically taken from classroom assignments, usually including teacher and student self-evaluations at regular (formal and informal) intervals. Finally, "on-demand" performance assessment tasks require students to respond to a complex scenario (e.g., word problems in mathematics or science, essays in language arts) within a fixed setting and time frame (e.g., 45 minutes in a classroom). I use the term "performance assessment" to include all of these (i.e., authentic assessments, portfolios, on-demand performance assessments), because they share the common features of complexity, constructed responses, knowledge integration, and authenticity.

The information in Table 1 suggests arguments supporting and challenging performance assessment. The merits of performance assessments versus traditional assessments have been debated elsewhere (Cizek, 1991; Hambleton & Murphy, 1992; cf. Wiggins, 1989). However, I will address the specific merits of performance versus traditional assessment approaches with respect to cultural, ethnic, and linguistic minorities.

**Performance Assessment and Minority Groups**

There are five key issues regarding performance assessment and minority groups:

1. Minority group scores on performance assessments.
2. Instructional alignment and performance assessments.
4. Performance assessment and language.
5. Test bias and performance assessment.
This article will examine the available data to address the use of performance assessments with minority groups. In particular, each issue will be addressed with respect to areas of consensus, areas of controversy, and areas yet to be studied.

**MINORITY GROUP SCORES ON PERFORMANCE ASSESSMENTS**

The first issue concerns the degree to which the minority/majority group achievement gaps noted on traditional assessments will be changed (or unchanged) by a shift to performance assessments. On the one hand, the gap could be smaller on performance assessments than on traditional measures because performance assessments are more likely to tap authentic reasoning, allowing for diverse and creative responses (instead of selecting the “one right answer”), and because performance assessments should be better aligned with instruction. On the other hand, the gap could be unchanged or larger using performance assessment methods because these demand more prior knowledge, greater transfer of knowledge across contexts, heavier language demands, and are more complex than traditional tests.

**Outcomes of Minority Groups on Performance Assessments**

The evidence regarding minority groups and performance assessment is remarkably limited, given the popularity of large-scale assessment programs and their importance for educational policy. However, Table 2 summarizes available studies addressing minority versus majority groups to illustrate the types of groups included in research, the academic domains assessed, the types of assessment, contrasts between minority groups and whites on these types of assessments, and how the majority/minority gaps vary across assessment methods.

The evidence in Table 2 suggests minority/majority group achievement gaps are often unchanged on performance assessments; some studies show smaller gaps, whereas others show larger gaps. One strong example of this conclusion is found in the research of Baxter, Sharelson, Herman, Brown, and Valadez (1993), which examined the performance of English-speaking Hispanic students on selected response (CTBS Mathematics) and performance assessment items. These items were developed to assess mathematical reasoning and knowledge as embodied in the National Council of Teachers of Mathematics (NCTM) standards, and included tasks such as estimation, reasoning, applied probability, and elaborated written communication of thinking and reasoning in problem-solving contexts. In addition, Baxter et al. (1993) added a condition in which students were prompted to use higher-order thinking and reasoning in developing their responses (in part because it was suspected that minority students may be less likely to apply spontaneously such skills to test situations than majority students). Furthermore, the assessment data were collected across two conditions: (a) traditional math instruction (emphasizing computation and rote skill development);
<table>
<thead>
<tr>
<th>Source</th>
<th>Minority Group&lt;br&gt;</th>
<th>Subject Matter</th>
<th>Assessment Types&lt;br&gt;</th>
<th>Minority vs. Majority contrast&lt;br&gt;</th>
<th>Comments&lt;br&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Badger, 1995</td>
<td>African American</td>
<td>Mathematics</td>
<td>MAP; On-demand</td>
<td>African American &lt; PA, MC</td>
<td>MC &gt; PA</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td></td>
<td>PA and MC</td>
<td>Asians = PA, MC</td>
<td>MC = PA</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td></td>
<td></td>
<td>Hispanics &lt; PA, MC</td>
<td>MC &gt; PA</td>
</tr>
<tr>
<td>Baxter et al., 1995</td>
<td>Hispanic</td>
<td>Mathematics</td>
<td>MC, PA</td>
<td>Hispanic &lt; MC, PA</td>
<td>MC &lt; PA</td>
</tr>
<tr>
<td>Bourque &amp; Garrison, 1991</td>
<td>American Indian</td>
<td>Mathematics</td>
<td>NAEP (MC and SA)</td>
<td>American Indian &lt; on MC, SA</td>
<td>MC = SA</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td>Asian = MC, SA</td>
<td>MC = SA</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td></td>
<td></td>
<td>Black &lt; MC, SA</td>
<td>MC ≤ SA</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td></td>
<td></td>
<td>Hispanic &lt; MC, SA</td>
<td>MC ≤ SA</td>
</tr>
<tr>
<td>Colwell &amp; Mitchell, 1993</td>
<td>Chapter One students</td>
<td>Language Arts</td>
<td>PA, MC, TJ</td>
<td>Chapter One &lt; PA, MC</td>
<td>MC &gt; PA, TJ</td>
</tr>
<tr>
<td></td>
<td>LeMahieu (cited in Bond, 1995)</td>
<td>African American</td>
<td>Writing</td>
<td>NAEP Essay</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Myerberg, 1996</td>
<td>African American</td>
<td>Mathematics</td>
<td>African Americans &lt; MC, SA, PA</td>
<td>MC, SA &gt; PA</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>Reading</td>
<td>MSPAP; MC, SA</td>
<td>Asians = MC, SA, PA</td>
<td>MC = SA</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>Written Language</td>
<td>PA</td>
<td>Hispanic &lt; MC, SA, PA</td>
<td>MC = SA</td>
</tr>
</tbody>
</table>

Note. 

aMinority groups were described using terms appearing in the original study.  
bMAP = Massachusetts Assessment Program; MC = multiple choice; MSPAP = Maryland School Performance Assessment Program; PA = on-demand performance assessment requiring elaborated written responses; SA = Short Answer; NAEP = National Assessment of Educational Progress; TJ = Teacher Judged portfolio score.  
c< group performed lower than white, majority group; = group performed equally well as white, majority group; > group performed higher than White, majority group.  
dContrasts describe the relative gap between minority and majority groups across assessment methods (e.g., MC, SA < PA means the gaps were equal for multiple choice and short answer methods, and both of those yielded smaller minority/majority differences than performance assessments).
and (b) hands-on instruction (emphasizing problem solving and integration of mathematical knowledge).

The findings surprised the researchers. Majority students outperformed minority students on traditional and performance assessments. Furthermore, this gap was not narrowed (and may have been widened) by prompts and hands-on instruction. These findings led Baxter et al. to conclude “contrary to expectation, the performance assessments did not have a salutary effect on the level of Latino students’ achievement. Rather...mean Latino-Anglo performance differences were larger on the performance measure than on the multiple choice achievement test” (1993, p. 213).

Three limitations of the available research merit discussion. First, most studies showing smaller ethnic group differences for performance assessments do not correct for differences in reliability between selected response and performance assessment methods. Studies that correct for reliability differences consistently report achievement gaps on performance assessments similar to or larger than gaps for selected response tests (Bond, 1995). Second, many studies examine data from the National Assessment of Educational Progress (NAEP), which typically compares multiple-choice scores to open-ended questions and essays (i.e., limited, on-demand performance assessments). Comprehensive evaluations, including an array of performance and traditional assessments (such as those used by Baxter et al., 1993), are rare.

Third, a growing body of research suggests strong specificity for performance assessment tasks (Baxter et al., 1993; Brennan & Johnson, 1995; Linn, Baker, & Dunbar, 1991). Because responses on one task do not generalize well to other tasks purportedly assessing the same subject matter, one cannot generalize results from one study (using one set of tasks) to performance assessments in general. Furthermore, differences may vary as a function of age and development. Longitudinal research (Sammons, 1995) suggests that majority/minority gaps on selected response and performance assessments may fluctuate over time, so that at times the gaps appear to be larger on performance assessment measures and at other times smaller on similar measures. It is not clear whether these differences are a measurement artifact or reflect genuine interaction between developmental progress and assessment methods.

Reviews of literature concur with the conclusion that performance assessments are unlikely to minimize majority/minority group differences in a meaningful way. For example, Bond states “the preliminary evidence for equity and adverse impact, defined simply as substantively different mean scores by [ethnic] subgroups, does not look good” (1995, p. 22). Winfield and Woodward conclude “the evidence [regarding performance assessments] collected thus far is inconclusive but suggestive of wider performance gaps between racial/ethnic groups” (1995, p. 14). Madaus concluded “the claims of proponents that authentic assessment will narrow the achievement gap simply lack supporting evidence” (1995, p. 36).
Areas of Consensus and Controversy

There are two areas of consensus evident on this issue. First, the evidence base is inadequate to draw strong generalizations. The data are limited in the range and scope of performance assessments investigated, and the specificity of performance assessment tasks further limits generalizations. Second, the available evidence suggests performance assessment tasks will not reduce minority/majority group achievement gaps. Indeed, the evidence to date suggests that, to the degree performance assessments affect minority/majority group achievement gaps, they are likely to make them larger.

Two controversies remain unresolved. The first is the degree to which contradictory research might change these conclusions. Given the limited research base, it is possible that more or newer research might change conclusions. The second is the degree to which majority/minority differences may be affected by development; it is not clear that differences among ethnic groups are stable across developmental states. Based upon the available data, it is unlikely that performance assessments alone will realize the hope of reducing minority group achievement inequities (Winkling & Bond, 1995).

INSTRUCTIONAL ALIGNMENT AND PERFORMANCE ASSESSMENT

Performance assessment aims to stimulate better alignment of curricula, instruction, and assessment. Hambleton and Murphy summarize this point by noting “If teaching to the test is inevitable... the test must be worth teaching to” (1992, p. 7). The corollary to this position is that minority group children should score better on performance assessments because these assessments more closely approximate what is taught, rather than emphasize prior knowledge and skills acquired outside of school.

Data Regarding Instructional Alignment and Minority Group Performance

It has already been noted that minority groups score no better on performance assessments than on traditional assessments. However, research uses large-scale, externally developed, and independently rated assessments. The performance of minorities on classroom-based assessments (i.e., those developed and used within classroom contexts) may produce different results. Research conducted on classroom-based assessments with high authenticity and complexity typically show smaller majority/minority group score differences (Newmann, Marks, & Gamoran, 1995; Newmann & Wehlage, 1995). Likewise, portfolio assessment procedures (i.e., samples of students’ classroom work collected over time) tend to show smaller minority/majority achievement gaps. However, this finding has been con-
tradicited in some studies (LeMahieu, cited in Bond, 1995) and may be a consequence of the poor reliability of classroom assessments. For example, Reckase (1995) estimates the mean reliability for portfolio ratings to be between .40 and .50; such low reliabilities severely attenuate minority/majority achievement gaps.

Unfortunately, research suggests but does not directly demonstrate that assessment outcomes affect instruction. Popular media accounts (e.g., The Washington Post, 1998) frequently claim poor test scores improve schools by spurring instructional changes. Consequently, studies finding large achievement gaps typically conclude that the cause of these gaps is inequitable access to instruction (Darling-Hammond, 1995; Kopriva, Lowrey, & Martois, 1994), whereas studies finding smaller achievement gaps typically conclude that equitable instruction is the cause of the narrower outcome (Newmann & Wehlage, 1995).

However, research does not support the notion that equitable teaching reduces or eliminates minority/majority student differences. Retrospective research shows low-scoring groups (e.g., English language learners) receive relatively few instructional opportunities to learn skills appearing in performance assessments (Kopriva, Lowrey, & Martois, 1994). However, prospective research (Herman, Klein, Wakai, & Heath, 1996) suggests few consistent learning opportunity differences between schools serving substantially different social classes and ethnic groups. Herman et al. concluded “...we did not find consistent differences across school types in students’ opportunities to learn” (1996, p. 22). This suggests that curriculum in general is poorly aligned to assessment tasks, but that learning opportunities are not the primary cause of high or low performance.

One study used a test-teach-test paradigm in which students were tested with performance assessments before and after instruction. The study (Baker, Niemi, Abedi, & Sato, 1993) found that ethnic group gains were larger for higher-scoring ethnic groups than for lower-scoring ethnic groups; students who initially had higher scores benefitted more from instruction than those with lower scores, resulting in larger gaps between Asians, whites, and Hispanics after equitable instruction. Although there may be differences in opportunity to learn among ethnic groups, it is neither clear that differential access to instruction creates majority/minority group differences, nor that equating learning opportunities for minority and majority groups will necessarily eliminate the gaps.

**Special Issue: Group vs. Individual Tasks**

A special problem emerges in the use of performance assessments in which students must work in groups to solve problems. Group tasks are used in some state assessment systems (e.g., California Learning Assessment System, Connecticut Common Core of Learning Assessment) to assess students’ proficiencies in academic content areas. However, group membership may affect students’ scores so that their score is influenced by their opportunities to interact with other students.
Some (e.g., Nueberger, 1993) argue that groups create greater instructional equity because heterogeneous work groups equalize learning opportunities across varying levels of performance, ethnicity, etc. Others note that access to high-performing students in such groups may itself be an equity issue. For example, students with limited prior knowledge generally perform worse when grouped with other students of the same knowledge levels, but score better when grouped with peers who have high degrees of prior knowledge (Lou, Abrami, Spence, Poulsen, Chambers, & d’Appollonia, 1996). Students with high abilities are apparently unaffected by group membership (Lou et al., 1996; Webb, Nemer, Chizhik, & Sugrue, 1997), and those of medium abilities may perform better when grouped with peers of equal skill (homogeneous groups) than when grouped with peers of different skill levels (heterogeneous groups; Lou et al., 1996; cf. Webb et al., 1997).

The use of group performance assessments then raises the issue of equity. If minority students are likely to be in the low ability/limited prior knowledge group and are more likely to attend schools where there are few high ability/prior knowledge students, they will be less likely to have the opportunity to improve performance assessment scores through group work. Interestingly, there is no evidence of a negative effect for high ability/prior knowledge students (i.e., being grouped with lower-ability students does not reduce high-ability students’ chances for success). Group performance assessments appeal to educators for their promise of equity, links to instruction emphasizing groups, and alignment with the goal of working with others. However, the desires to achieve the benefits of collaborative work groups in performance assessments must be balanced with the need to ensure equity across testing conditions for all (low, medium, and high ability) students.

Areas of Consensus and Controversy

The central premise of standards-based reform is that richer, more complex learning assessments will induce more effective and complex instruction for all students. There are few data to support this contention, although studies of effective school reform suggest that effective schools align instructional and administrative supports to enhance outcomes on authentic learning tasks (Lee, Smith, & Croninger, 1995; Newmann & Wehlage, 1995). Baker, O’Neil, and Linn concluded “...claims that the new forms of [performance] assessment will have positive effects on instruction and student learning need to be supported by evidence” (1993, p. 1216).

Data regarding equity of assessment outcomes for assessment aligned with instruction are controversial. Whereas classroom-based studies tend to show narrower majority/minority achievement gaps, studies using national- or district-based performance assessments show similar or larger gaps (Simmons, 1993). Whether the narrower gap noted for classroom-based studies is a function of instructional alignment (i.e., assessing students on what was taught reduces gaps among children), or a function of the poor reliability of classroom-based ratings, is not yet known. The available evidence points to poor reliability, not instructional
PERFORMANCE ASSESSMENT AND DIVERSITY

Equity, as the cause of smaller minority/majority achievement gaps on classroom-based performance assessments. Studies controlling instructional opportunities do not eliminate (and may even increase) performance assessment differences among ethnic groups. Furthermore, group work in performance assessments is controversial in that it promises greater equity through heterogeneous problem solving and peer interactions, yet may inadvertently reduce low-ability students' chances for success if groups do not include at least one high-ability member.

PERFORMANCE STANDARDS AND MINORITIES

All but one state in the U.S. (Iowa) has state academic standards (National Center for Educational Outcomes, 1999). There are three types of academic standards:

1. Content standards, which state in broad terms what students are to know (e.g., students in Wisconsin will apply their knowledge of the nature, grammar, and variations of American English).

2. Performance standards, which detail what students should be able to do at a given point in their educational careers (e.g., by the end of Grade 4, students will...develop their vocabulary of words, phrases, and idioms as a means of communication).

3. Proficiency standards, which specify how well students must do on an assessment to be judged as meeting (i.e., being proficient in) the performance standard (e.g., a rubric to judge essays as minimal, basic, proficient, or advanced levels of proficiency).

Although performance assessments may be used independently of standards, they are often used to assess proficiency. As districts and states specify what they want students to know (content standards) and do (performance standards), they define performance assessments to determine whether students meet proficiency standards. Thus it is not possible to discuss performance assessments without also discussing performance standards (Messick, 1995).

Assessment Purpose and Acceptability

Whereas some critics view the process of standard setting as an exercise in cultural hegemony (Merelman, 1996), others view standards as necessary for eliminating ethnic group inequities (Roeber, 1995). These antithetical views assume different purposes for assessment. When the primary purpose of assessment is assumed to have low stakes (e.g., school/district accountability, routine classroom assessment), assessment equity advocates are more likely to view performance assessments as useful in reducing inequities (Darling-Hammond, 1995; Madaus, 1995; Winfield, 1995; Winkling & Bond, 1995). However, when the primary purpose of assessment has relatively high stakes for individuals (e.g., student graduation/re-
tention/promotion) and educators (e.g., allocation of funds), equity advocates dispute the value of assessments. Some advocates view high-stakes performance assessment as a mechanism for denying minority students learning opportunities, because low-scoring students are retained, denied graduation, or lose learning benefits (Garcia, 1986; Obiakor, 1993). In contrast, other advocates view high-stakes assessments as necessary to change the status quo, arguing that minorities must be held to the same standard as majority students and that educators must be held accountable for eliminating minority/majority achievement gaps (Holmes, 1986; Roeber, 1995).

**Opportunity to Learn/Instructional Delivery Standards**

Two efforts aim to ease the tension between high standards and minority group failure. First, educators can reduce adverse impact on minorities by providing multiple methods of assessment, and allowing students and their families some control over assessment methods (Linn, Baker, & Dunbar, 1991; Rothman, 1993). Second, educators should adopt “opportunity to learn” standards (Glaser, 1997). These standards (also called “delivery” standards) require schools to provide all students opportunities to learn and master complex knowledge and skills. Advocates argue that opportunities to learn must be equalized before common standards for promotion, graduation, and the like can be applied fairly (Shepard, 1992). These advocates assume that (a) educational opportunities are not equally distributed across majority and minority groups; and (b) the majority/minority achievement gaps are caused by opportunity differences (and will thus be eliminated when opportunities are equalized).

There is ample evidence for the assumption that educational opportunities are unequally distributed across ethnic and social classes. Classroom activities, teacher philosophies, and fiscal supports vary widely across school districts and are largely associated with the socioeconomic status (SES) of school communities (Kozol, 1992; Oakes, 1985, 1990). Because SES and ethnicity are confounded in American society, differences in learning opportunities by social class lead to differences in learning opportunity by ethnic and linguistic groups. Although funding and other aspects of learning opportunity undoubtedly vary between districts and schools (Kozol, 1992), it is not clear that instructional opportunities vary substantially within classrooms and schools by ethnicity (cf. Oakes, 1990).

Evidence supporting the second assumption (i.e., that learning opportunities cause differences in academic achievement) is not provided. Advocates assume that differences in classroom learning opportunities—not genetic, early childhood, or nonschool environmental factors—lead to different achievement outcomes. This is an ideological, not evidential, position. Although some studies show similar outcomes in response to common instruction (e.g., Newmann & Wehlage, 1995), other (more carefully controlled) studies do not (Baker, Niemi, Abedi, & Sato, 1993; Baxter et al., 1993). Likewise, learning opportunities, as
currently measured, are not associated with differences in performance assessment outcomes (Herman et al., 1996). There is a substantial history in educational research demonstrating the difficulty of long-term change in learning trajectories (Spitz, 1987). Furthermore, ethnic group differences found within classrooms, schools, and school districts are not readily explained by SES differences, but because the achievement gap exists, advocates cite opportunity to learn as the cause of (Darling-Hammond, 1995) or the solution to (Roeber, 1995) minority/majority achievement gaps.

It is politically unacceptable to suggest that equating learning opportunities might not eliminate minority/majority achievement gaps. For example, the U.S. President (Executive Office of the President, 1996) rejected IQ and family circumstances as determinants of success and demanded that schools be held responsible for educational outcomes (regardless of student IQ and family circumstances). This might be logically equivalent to demanding that astronomers take responsibility for celestial noncompliance with a geocentric model of the solar system, or demanding that physicians take responsibility for ethnic group differences in hypertension.

The demand for common outcomes regardless of individual and class differences has gone largely unchallenged (cf. Braden, 1997; Coffman, 1993). Those who believe classroom learning opportunities are the cause of, or solution to, the minority/majority achievement gap have characterized any other position as succumbing to lower standards for minority groups. A “middle ground” position—making no assumptions about how groups will fare if provided similar learning opportunities—is characterized as a “self-fulfilling prophecy dooming significant proportions of students to educational and social failure” (Roeber, 1995, p. 266).

Areas of Consensus and Controversy

Few dispute the assumption that low-stakes use of common standards will benefit minority and majority group students. By calling attention to poor performance, it is assumed (but not empirically demonstrated) that schools will improve instruction to improve student test scores. Many dispute the value of performance standards for high-stakes uses (e.g., promotion, graduation). A primary concern is that minority students may be denied equal learning opportunities and are therefore unfairly punished by high-stakes assessments. Two popular responses to this concern are (a) to provide minorities choices in assessments; and (b) to develop learning opportunity standards. Educational policy makers apparently believe that eliminating learning opportunity differences will eliminate learning outcome differences, despite substantial contradictory evidence. Political pressure and the characterization of other alternatives as a self-fulfilling prophecy for failure have encouraged the belief that equalizing opportunities to learn will equalize achievement outcomes, and have discouraged discourse holding different or “no assumptions” positions.
PERFORMANCE ASSESSMENTS AND LANGUAGE

One of the critical issues surrounding performance assessments and ethnic groups is the portability of performance assessments across languages. Regulations clarifying the Improving America’s Schools Act require states to assess student progress using linguistically appropriate assessments for English Language Learners (ELLs) (Aguirre-Muñoz & Baker, 1997). Three questions emerge in response to this mandate: (a) Should performance assessments be translated (i.e., should educators demand that all students speak, read, and write in English)? (b) If so, can they be translated effectively? and (c) If so, who should be offered translated assessments?

Should Performance Assessments Be Translated?

The first issue is one of policy, with some insisting that all students must acquire mastery of English, and should therefore be required to demonstrate such mastery as evidence of educational success. Others demand pluralistic standards to recognize competency in other languages (Gonzales, 1993). Such either/or thinking polarizes positions; a middle-ground strategy argues that students should be given opportunities to demonstrate their competence in languages in addition to English (see “Guidelines for Equitable Assessment” by the Diversity and Equity in Assessment Network, 1993; cited in Neill, 1995). Ultimately, this issue is political and not scientific, although some argue such decisions should be included within the scientific assessment arena as an instance of consequential validity (Messick, 1995). Given federal regulatory directives, the answer to the first question appears to be “yes.”

Can Performance Assessments Be Translated Effectively?

The plethora of position statements and guides for how to translate performance assessments into languages other than English (Geisinger & Carlson, 1992; Koelsch, Estrin, & Farr, 1995; Neill, 1995) presume that successful translation is possible. But is this presumption accurate? The primary vehicle for evaluating the success of a translation draws on item response theories (see also the next section of this article). These technologies require many test takers and many items (Sireeci, 1997). This is a significant challenge for performance assessments, because the items (i.e., tasks) are typically long and complex, and therefore few are given in an assessment. Also, complex scoring demands (i.e., raters must read and rate responses) typically limit the number of students taking a test. Two forms of evidence are consistently rejected as evidence of problematic assessments: (a) differential pass/fail rates (i.e., it is not assumed that a successful translation necessarily eliminates between-group differences); and (b) expert judgments. Although expert judgments are useful for initial item development and avoiding inappropriate language uses, they rarely discriminate effective from ineffective translations. To date, there are far more posi-
tion statements and guides than evidence (Hambleton & Murphy, 1992). Performance assessments are expensive to develop (Cizek, 1991; Rothman, 1993) and interpret (e.g., translations, large numbers of items, large numbers of non-English-speaking test-takers), which may limit the available evidence. The limited evidence to date (Budgell, Raju, & Quartetti, 1995) suggests sophisticated item response methods can detect successfully translation problems in performance assessments, and that translation problems may not be particularly prevalent in well-executed assessment translations.

Who Should Receive Translated Assessments?

The primary issue in addressing the target audience for translated performance assessments is economic. Some school districts enroll children speaking more than 40 different languages; translation of performance assessments into all of these languages is not economically viable. Therefore, part of the question for “Who gets translations?” is “Which languages are sufficiently prevalent to justify translations?” It is not economically feasible to offer all students an assessment in their native language, yet districts and states must balance equitable access with economic reality.

Despite mandates to offer translated assessments, and some efforts to define “best practices” for such translations, researchers have paid scant attention to the question “How should students be selected/targeted for translations?” (Aguirre-Muñoz & Baker, 1997). Typically, students are screened in English (their second language, or L2), and are offered a translated assessment in their primary language (L1) if they perform poorly. The irony is that there is no direct assessment of L1 to determine whether they have the requisite language skills for the translated assessment. Thus, variability in L1 proficiency (which is assumed to be substantial in many ELLs) presents a significant challenge to interpretation of performance assessment results among non-English-speaking groups.

Areas of Consensus and Controversy

There is consensus that students with limited English proficiency should be provided performance assessments in their native language in areas other than language arts. There is controversy regarding whether language arts assessments should be translated into languages other than English. This issue is political, not scientific. There is consensus regarding methods to determine the accuracy of linguistic translations of performance assessments; however, these methods may be impractical for most performance assessments because of their complexity. Given the expense of developing and scoring performance assessments and the relatively small numbers of test takers and items, it is unlikely that language translations will be adequately developed and evaluated before use. Finally, current practices for targeting students fail to measure separately language proficiency in L1 and L2, confounding translation issues with native and secondary language proficiency.
One issue looms over all assessments with ethnic minorities: test bias. In this discussion, bias is assumed to mean systematic mismeasurement of ethnic minority individuals. Such systematic mismeasurement would include consistently underestimating achievements, consistently overestimating achievements, or less accurately estimating achievements (i.e., more error) relative to majority populations (Jensen, 1980; Reynolds, 1998). Although scientists reject differences in passing and failing rates as evidence of bias, lower pass rates on performance assessments have been cited as evidence of test bias by the media and educational assessment critics (Frazier & Nakashima, 1998).

Methods for Detecting Bias in Performance Assessments

The essential question all methods seek to answer is “Does the assessment function differently within different groups?” These methods range from collecting evidence at item levels (using differential item functioning [DIF] methods), detection of rater bias (i.e., do raters systematically rate responses from some groups different than responses from other groups?), differential scale functioning (e.g., different reliability or validity statistics across groups), and predictive validity (i.e., does group membership influence the accuracy of assessment decisions with respect to an external criterion?). A final question that is rarely raised with respect to cognitive tests, but is essential for achievement tests, is “Are the domains sampled by this test fair and appropriate for all groups assessed?” Domain representation (“content validity”) is an important concern for all achievement tests, including performance assessments.

Evidence for Bias in Performance Assessments

DIF Studies. Although some dispute the value of DIF approaches for detecting bias (Gonzalez-Tamayo, 1988), there are many who champion the use of DIF to detect bias in all forms of assessment, including performance assessments (Dorans & Potenza, 1994). Although most DIF studies of items with multiple-response categories (i.e., polytomous items) have focused on the constructed response items on the Mathematics portion of the Scholastic Assessment Test (Lawrence, Lyo, & Feigenbaum, 1995; Lyu, Dorans, & Ramsay, 1995), some have investigated richer performance assessments (Ferrara & Walker, 1990). Most studies find limited or no DIF on items for major ethnic groups, and no consistent patterns favoring one group over another. For example, Ferrara and Walker (1990) found small DIF (i.e., item bias) in narrative essays favoring minority groups, and small DIF in explanatory essays favoring majority groups. However, they also noted that DIF is difficult to use with performance assessments due to model fitting assumptions and parameters. They noted “significant DIF statistics [may] reflect measurement, calibration, or sampling error rather than a real disparity in item functioning” (Ferrara &
Walker, 1990, p. 25). Ironically, efforts to equate respondents’ educational background in DIF analyses yield higher DIF which favors lower-scoring minority groups, suggesting that either these methods are methodologically unsound, or that DIF favors low-scoring ethnic groups (but these effects are attenuated by educational background variables) (Schmitt & Crone, 1991).

**Rater Bias.** Detection of rater bias is quite limited. Although there are many studies (using G or generalizability theory) showing that well-trained raters are generally quite consistent in rating student responses (Brennan & Johnson, 1995; Linn, Burton, DeStefano, & Hanson, 1996), there are few studies addressing the interaction of raters with respondents’ ethnicity (i.e., rater bias). It is possible that rater bias may influence the accuracy of performance assessments across ethnic groups. First, performance assessments are more complex and are more likely to engender disagreements among raters than objective assessments (Baker & O’Neill, 1995). This may be especially true for minorities, as their responses may be less likely to conform to established (i.e., dominant culture) scoring rubrics and thus generate more scoring disagreements (i.e., rating inaccuracy). Second, teachers’ command of subject matter is positively related to their accuracy and consistency as raters (Baker, Linn, & Abedi, 1993). If more knowledgeable teachers are more likely to work in affluent schools enrolling few minorities, then it follows that minority student responses are more likely to be rated by less knowledgeable, and therefore less accurate, raters. Baker and O’Neill (1995) speculated that raters of performance assessments may rate members of their own ethnic group higher than members of other ethnic groups. Because minorities are underrepresented in the teaching and rater ranks, it follows that majority group raters might underestimate scores of minority group children.

The available research (Howell, Bigelow, Moore, & Evoy, 1993; LeMahieu, cited in Bond, 1995) contradicts these speculations. LeMahieu found that ratings of African American students’ language arts portfolios were not associated with the race of the rater. Furthermore, LeMahieu found teachers familiar with students were likely to rate portfolios higher than when unfamiliar with students, suggesting halo effects. Howell et al. (1993) found evidence of rater bias, but bias in favor of Hispanic minorities. That is, raters gave higher ratings to identical passages when the author of the passage was identified as a lower SES Hispanic than when the passage author was identified as upper SES white. Note that higher ratings do not necessarily help minority students; rather, higher ratings imply that raters have lower standards for minority students than for majority students.

**Differential Scale Functioning.** Differential scale functioning is rarely addressed in performance assessment research. Because performance assessment scales rarely contain many items (tasks), traditional measures of reliability and validity are less likely to be useful in evaluating scale characteristics. Furthermore, performance assessments have much higher item-specificity than traditional tests (Baxter et al., 1993; Brennan & Johnson, 1995; Linn et al., 1996), making it difficult to justify using multiple tasks to form a common scale. Therefore, the types of informa-
tion often available in studies of bias within intelligence tests (e.g., item difficulties, delta decrements, differential reliabilities in scales, correlations among items and scales) are not typically available for evaluation of bias within performance assessments. The available evidence suggests that performance assessments are often less reliable than traditional tests (Bond, 1995). Furthermore, when majority/minority achievement differences are corrected for poor reliability, gaps generally get larger rather than smaller (Bond, 1995). However, performance assessments can achieve substantial reliability when used within ethnically diverse, culturally disadvantaged groups (Plucker, Callahan, & Tomchin, 1996).

DeVaney and Franks (1995) evaluated the construct validity of performance assessments by factor analyzing multiple mathematics achievement indicators. They found that performance assessments were substantially independent of other indicators of mathematics success (i.e., Stanford Achievement Test scores and teacher grades) within African American elementary students. The lack of congruence among multiple indicators is distressing, but quite common in performance assessment research. Other analyses of performance assessments in predominately majority samples (Braden, Elliott, & Kratochwill, 1997) also show little overlap or agreement between performance assessments, grades, and traditional achievement measures. Myerberg (1996) also found moderate to low correlations among multiple achievement indicators, although correlations among these variables may be higher for minority groups.

Predictive Validity. One alternative that is rarely considered is that an assessment may be biased by being more (not less) accurate for a minority group than for a majority group. Hambleton and Murphy (1992) suggest this may be the case for performance assessments. This raises an interesting question: Which measure should be used as the predictor, and which as the criterion, for judging predictive validity? Teacher grades are far less reliable than other measures of academic achievement, and other measures of academic achievement are narrower and frequently less aligned with standards than performance assessments. Perhaps in part because of this issue, performance assessments in K–12 education have not been evaluated for predictive bias. Studies of performance assessments in vocational settings (e.g., in-basket exercises and other assessment center methods) suggest that intercept bias (in which the criterion performance of low-scoring groups is overpredicted, and the criterion performance of high-scoring groups is underpredicted) is common (Gottfredson, 1994).

A second problem with predictive validity studies of performance assessments is criterion contamination. In a traditional study of predictive bias, how a person does on the test should not affect how the person does on the criterion. This is usually accomplished by making criterion judgments without knowledge of the test results. However, because performance assessments are often used for educational decision making, prediction of graduation, grade point average, or other criteria may be contaminated with the assessment.
Performance assessments based on multiple intelligences theory have shown acceptable reliability, but have poor predictive validity for identifying gifted students from ethnically diverse, low-income backgrounds (Plucker et al., 1996). I could find no predictive validity studies more relevant to common performance assessment uses (e.g., high school graduation) that were not confounded by criterion contamination. The use of performance assessments as a criterion (e.g., for grade promotion, graduation) precedes support for their classification as an unbiased predictor of educational performance.

**Domain Representation/Content Validity.** Many writers express concerns regarding the cultural narrowness of performance assessments, which emphasizes dominant cultural knowledge and processes to the exclusion of minority perspectives (Darling-Hammond, 1995; Neill, 1995). Methods for ensuring cultural sensitivity through panels of ethnically diverse educators who teach ethnically diverse students can be helpful in developing items, but these panels are not reliable for identifying inappropriate content representation after item administration (Ferrara & Walker, 1990). Scientific methods for domain sampling are poorly conceptualized and developed; currently, test makers ensure ethnic representation as items are developed to ensure items contain culturally and linguistically sensitive content (Koelsch, Estrin, & Farr, 1995).

**Areas of Consensus and Controversy**

Critics and the media often presume that, to the degree bias affects performance assessments, it is likely to underestimate the achievements of minority students (Frazier & Nakashima, 1998). There is no evidence to support this contention. DIF studies show small effects routinely favoring lower-scoring minority groups. The limited available research shows raters rate minority students’ responses equal or higher than those of majority students. It is possible that raters are less accurate when rating the responses of minority students (i.e., differential accuracy bias), but there is no evidence directly addressing this issue. There is no evidence to suggest performance assessments function differently for groups within the performance assessment scale (i.e., differential reliability), nor with respect to other measures (i.e., differential validity). However, it is clear that performance assessments measure something not shared by other measures of achievement, which means they are not highly related to teacher grades or traditional achievement tests. Finally, there is virtually no evidence to address whether use of performance assessments yields differential prediction within K–12 educational applications. Despite their increasing use in high-stakes assessments (e.g., decisions regarding promotion, graduation), there is virtually no evidence to support performance assessment use as an unbiased predictor. (Of course, there is also no evidence to suggest they are a biased predictor.)
Ironically, when performance assessments show smaller minority/majority achievement differences, investigators often presume the smaller difference is evidence that performance assessments are better measures of minority student achievements (Badger, 1995; Myerberg, 1996). Conversely, larger achievement gaps are taken as proof of bias in tests or instruction. Methods for detecting bias independent of group means must be developed and applied to performance assessments.

**DISCUSSION**

There are many issues surrounding performance assessments and ethnic/linguistic diversity. The available literature suggests that although performance assessments may offer more meaningful assessment opportunities than traditional selected-response assessments, performance assessments do not yield smaller differences among ethnic groups. Although ethnic groups may score more similarly on within-classroom performance assessments than on standardized, externally developed performance assessments, smaller differences may be due to poor reliability rather than equitable outcomes.

Experts disagree about the relative benefits of performance assessments for minority students. Whereas low-stakes performance assessments are generally popular, high-stakes performance assessments are controversial. High-stakes advocates argue educators who demand equal outcomes for all ethnic and linguistic groups will get equal outcomes (i.e., expectancy theory is sufficient to eliminate ethnic and linguistic performance differences). High-stakes assessment critics argue educators must first provide ethnic and linguistic groups equal learning opportunities before using assessments for high-stakes outcomes. An agnostic position on group differences (e.g., neither expectancy theory nor in-school learning opportunities are sufficient to eliminate differences between groups) is politically unacceptable. Advocates promote all-or-none thinking (e.g., if educators do not demand equal outcomes for all groups, then they are perpetuating outcome inequities) that inhibits thoughtful dialogue.

Despite some significant practical challenges (e.g., costs of developing, scoring, translating) and some theoretical challenges (e.g., generalizability of performance across raters, tasks, settings, time), performance assessments are an appealing complement to selected-response tests. Their use with ethnic and linguistic minorities shares some of the problems associated with traditional tests (e.g., cultural content, demand for prior knowledge, sensitivity to instructional inequities), and some problems unique to performance assessments (e.g., rater bias, higher cost, unique variance associated with performance assessment methods). Interestingly, performance assessments appear immune to one of the most popular criticisms of traditional tests: arbitrary content. Because performance assessments are closely linked to instructional practices and curricular standards, and more closely resemble instructional practices than selected response (i.e., traditional) tests, educators may be more inclined to accept between-group differences on performance assessments as genuine.
REFERENCES


Holmes, B. J. (1986). Do not buy the conventional wisdom: Minority teachers can pass the tests. *Journal of Negro Education, 55*, 335-346.


Action Editor: Randy W. Kamphaus
Acceptance Date: May 12, 1999