Ethnic-Based Equity in Teacher Judgment of Student Achievement on a Language and Literacy Curriculum-Embedded Performance Assessment for Children in Grade One

By Dorinda J. Gallant & James L Moore, III

Dorinda J. Gallant is an assistant professor in quantitative research, evaluation, and measurement in the College of Education at The Ohio State University, Columbus, Ohio. James L. Moore III is an associate professor in the College of Education and Human Ecology, coordinator of the School Counseling Program, and the inaugural director of the Todd Anthony Bell National Research Center on the African American Male at The Ohio State University, Columbus, Ohio.

In today's American educational system, testing has increasingly become the preferred educational practice for measuring students' academic achievement. It is often seen as the most suitable means to raise academic achievement for students, by insisting on educational accountability among educators—teachers, school counselors, and administrators (Moore, 2003). Given the steady decline of school achievement among students in the United States, greater interest in public education now exists. Such interest extends beyond the walls of professional educators, district and building administrators, local school boards, and state education departments. Increasingly, business and political

leaders are involving themselves in educational policy, in turn, to improve school achievement in America's public schools (Usdan, 2006). These leaders' involvement in public education evolved out of their concern that the United States may no longer be producing students who are able to compete in today's global economy (Friedman, 2005).

Because public education is commonly seen as the "driving force" of the country's economic prosperity, public education is an important topic of debate (Moore, 2003; Southern Education Foundation, 1995, 1999, 2000a, 2000b), and desperate measures are frequently offered to improve public education for *all* students (Moore, 2003), especially low-income students, students with disabilities, and students of color. Too often, students with these particular backgrounds find themselves in failing schools or educational systems that do not give them the skills needed to succeed in the "new" global workplace (Kozol, 1991; Sears, 2002). There are a number of variables that interact to impede school outcomes for these students. They range from chronic poverty, low teacher expectations, poor parental involvement, and inadequate school funding and resources (Flowers, Milner, & Moore, 2003; Ford & Moore, 2004; Kozol, 1991; Sears, 2002). Nevertheless, based on national statistics, it is quite clear that these students are being left behind in these schools, and society offers them few assurances that they will be able to compete in today's global economy. Therefore, the demands from business and political leaders to educators—teachers, school counselors, and administrators—to leave no child or group behind are seen as warranted. It is increasingly apparent, at least from these leaders, that public schools will not be able to reform themselves without assistance or force from outside entities (Usdan, 2006). This belief, of course, mirrors earlier calls for action to improve the quality of education in the United States (Cobia & Henderson, 2007).

Arguably, no other time in public education has generated as much attention on the quality of America's educational system than the seminal report *A Nation At Risk* (National Commission on Excellence in Education, 1983). This document revealed that America is no longer leading the world in educating its citizens. The report further indicated that students in the United States lagged drastically behind their student counterparts in other countries (National Commission on Excellence in Education, 1983). In recognition of these educational trends, special efforts—from the business and political communities—are frequently given to reform or transform public education (Usdan, 2006). As an example of this focus, the National Governors' Association (1986) endorsed in *Time for results: The governors' 1991 report on education* high educational standards and adequate measures to determine whether students were meeting the identified standards. Consequentially, these standards evolved into "high stakes" testing (National Council on Education Standards and Testing, 1992). Testing, in other words, is seen "as a viable measure of academic achievement and a suitable means of accountability" (Cobia & Henderson, 2007, p. 35), which is the primary basis of the current No Child Left Behind Act of 2001 (U.S. Department of Education, 2001, 2004). Further, this federal mandate em-

phasizes stronger school accountability through testing. It represents the "driving force" of public education throughout the United States.

In many school systems, the *No Child Left Behind* legislation has become the primary focus of schools but, at times, at the expense of student learning. Many educators are beginning to teach to the test to ensure that students achieve passing scores on statewide testing. Because the law requires school districts to report students' test scores, educators, particularly in urban settings, have grown hypersensitive about testing their students. They have grown to see these tests as being punitive to students and schools rather than helpful in improving deficits in students' learning. Moreover, many educators—teachers, school counselors, and administrators—are growing overwhelmed by the pressures placed on them as a result of the *ongoing* testing requirements in public schools, as well as the demands to meet certain state benchmarks. It is widely accepted that many students of color, such as African Americans, do not perform well on educational tests (Ferguson, 1998; Ford, 1996). Aligned with this notion, Ford (1996) asserts:

> Arguments against using standardized tests with Black students have a long history. These arguments have proliferated in recent years on the grounds that ethnically and culturally diverse students are assessed by tests that do not effectively measure their intelligence and achievement. Specifically, because of the life experiences and educational opportunities of minorities and White students vary considerably, the reliability of traditional standardized tests may be questionable for Black and other minority students. (p. 55)

A significant characteristic of any test is the degree to which it is valid and reliable (Ford, 1996; Ford, Grantham, & Bailey, 1999). Therefore, issues related to equity and fairness in educational testing and assessment, including test development and selection, test administration and scoring, and test reporting and interpretation of results, are important aspects of the testing process. Fairness in testing is described as: (a) a lack of bias in consistent score meanings across examinee subgroups, (b) equitable treatment of all examinees in the testing process, (c) equality in testing outcomes for examinee subgroups, and (d) equity in opportunities for examinees to learn tested material (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). In this study, fairness in testing is defined as a lack of bias in consistent score meaning across examinee subgroups.

Over the years, equity and fairness issues in testing have primarily focused on students' achievement on standardized tests. There are several studies directly related to the Black-White achievement gap (e.g., Ferguson, 1998; Lee, 2002). Many of these research investigations explored and discussed the role of test bias, heredity, and family background in the achievement gap; reasons for changes in the achievement gap over time; educational, psychological, and cultural explanations for the achievement gap; and educational and economic consequences of the achievement gap (see Jencks & Phillips, 1998). The cohesiveness of these studies provides a comprehensive understanding of contributing factors and consequences

of the Black-White achievement gap. In an effort to narrow ethnic and racial achievement gaps, several states demonstrated success through major educational reforms. These educational initiatives included, but not limited to, smaller class sizes, voucher programs, and comprehensive effective schooling models (see Chubb & Loveless, 2002). It is quite likely that the success of these reforms could provide examples for other local and state assessment programs.

Educational reforms in testing and assessment have prompted the use of performance assessments as an alternative to traditional standardized tests in local and state assessment programs. The potential of performance assessment to bridge the achievement gap between ethnic and racial groups has been received with mixed reactions. Darling-Hammond (1995) suggested that students' performance on authentic assessments provide teachers, administrators, and policy makers with a broader perception of students' academic skills and ability which is important in decreasing bias and giving educators the ability to make more informed decisions about students in terms of tracking, retention, and allocation of rewards in schools compared to traditional standardized tests. Meisels, Dorfman, and Steele (1995) also argue that performance assessment permits teachers to focus on teaching higher order academic skills rather than focusing on lower-level skills. However, performance assessment can potentially broaden achievement gaps for students of different ethnic groups, and as a result lead to less access and opportunities of different ethnic groups (Baker & O'Neil, 1995).

On another note, very little empirical research has been conducted on equity issues in the use of performance assessments. Perry and Meisels (1996) extended issues of technical adequacy of teachers' judgments to include equity. In their review of the literature, support was not found for gender or ability biases for teachers, as a group. However, they imply that the existence of individual biases may make it difficult for teachers to judge the academic performance of children. Perry and Meisels suggest additional research to explore the potential impact of linguistic and cultural differences on teachers' judgments. Similarly, Flowers et al. (2003) found, when investigating African American high school seniors' educational aspirations, that the students' perceptions of their high school teacher's expectations of their educational future had a significant impact on educational aspirations. This further suggests that teachers' behaviors, whether verbal or nonverbal, can influence students' school outcomes.

The purpose of this study was to determine the extent to which ethnic-based differences exist in teacher ratings of African American students and White students on the language and literacy domain of a curriculum-embedded performance assessment for children in grade 1. Further, the study extended previous research on performance assessments to focus on issues related to equity in teacher ratings on a curriculum-embedded performance assessment for young children in an urban school district. The research question of interest was: "To what extent do teacher ratings of students' performance on the language and literacy domain of a curriculum-embedded performance assessment differ based on students' ethnicity?"

Method

Participants

This study included the data files of 1,761 first-grade African American (n=1,442) and White (n=319) students in an urban school district in the southeastern part of the United States. To be included in this study, students' records needed to be complete for each indicator on the language and literacy domain. Thus, only African American and White students who received ratings on each item were included in this study. The demographic characteristics of the students were 82% African American, 52% male, 71% receiving free or reduced-price lunch, and 12% with an individualized education program (IEP). The demographic characteristics of the district's first-grade teachers (n = 138) with valid information (i.e., the inclusion of demographic information) were 52% White, 47% African American, 1% Asian, all female, and 70% with a bachelor's degree plus 18 hours and above.

Measures

The data used in this study came from the 2002 spring administration of the language and literacy domain of a state-wide curriculum-embedded performance assessment for students in first grade. The curriculum-embedded performance assessment is based on the personal and social development, language and literacy, and mathematical thinking developmental guidelines and checklists of the Work Sampling System™ (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994).

The Work Sampling System. The WSS is a curriculum-embedded performance assessment designed to "assess and document children's skills, knowledge, behavior, and accomplishments as displayed across a wide variety of classroom domains and as performed on multiple occasions" (Meisels, 1993, p. 36) for children in preschool through fifth grade. The WSS consists of three components: (1) developmental guidelines and checklists, (2) portfolios, and (3) summary reports. Developmental guidelines and checklists are designed to "assist teachers in observing and documenting individual children's growth and progress" (Meisels, 1993, p. 36) three times a year in seven curriculum areas: (a) personal and social development, (b) language and literacy, (c) mathematical thinking, (d) scientific thinking, (e) social studies, (f) the arts, and (g) physical development. The curriculum areas are further divided into functional components and performance indicators.

Developmental guidelines provide a set of criteria based on national, state, and local curriculum standards for observing and evaluating each checklist performance indicator. Thus, the guidelines enhance the observation process and increase the accuracy of teachers' evaluations. The other components of the WSS are portfolios and summary reports. Portfolios are "a purposeful collection of children's work that illustrates their efforts, progress, and achievements and potentially provides a rich documentation of each child's experience throughout the year" (Meisels, 1993, p. 37). Summary reports consist of "a brief summary of the child's classroom

performance and is based on teacher observations and on records that teachers keep as part of the Work Sampling System" (Meisels, 1993, p. 38). The reports translate information obtained from developmental checklists and portfolios into a document that is easily read and understood by parents, teachers, and administrators. Summary reports are completed three times a year.

Reliability and validity evidence for the WSS have been conducted by Meisels and associates. The internal consistency of teacher ratings has ranged from .87 to .94 on the art and fine motor (12 items), movement and gross motor (11 items), concept and number (15 items), language and literacy (17 items), and personal/social development (14 items) domains of the WSS checklists (Meisels, Liaw, Dorfman, & Nelson, 1995). Concurrent validity evidence has shown that approximately 75% of the correlations between teacher ratings and a psycho-educational measure ranged between .50 and .75 (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). Furthermore, 92% of the correlations between teacher ratings and student achievement were between the moderate to high range (Meisels et al., 2001). Predictive validity evidence has shown correlations ranging from .67 to .76 between teacher ratings and students' academic performance on psycho-educational measures (Meisels et al., 1995).

For this study, the domain of interest was the language and literacy domain, and it was selected because of national interest in the performance of students in the area of English language arts and reading. The language and literacy domain consisted of 12 indicators in the areas of listening, speaking, literature and reading, and writing. Example indicators required students to demonstrate that they could follow directions that involved a series of actions, read and comprehend fiction and non-fiction text, and write to communicate ideas. Using the developmental guidelines and checklist, teachers rated students' performance on each of the 12 indicators on an ordinal scale ranging from 1 to 3, where 1=*not yet*, 2=*in process*, and 3=*proficient*, three times (i.e., fall, winter, and spring) during the 2001-2002 academic year. Descriptors for the ratings are as follows:

> 1. *Not Yet* indicates that the skill, knowledge, or behavior has not been demonstrated.
>
> 2. *In Process* indicates that the skill, knowledge, or behavior is emergent, and is not demonstrated consistently.
>
> *3. Proficient* indicates that the skill, knowledge, or behavior is firmly within the child's range of performance. (Dichtelmiller, Jablon, Meisels, Marsden, & Dorfman, 1998, p. 41)

Statistical Analysis

The ordinal logistic regression procedure (Zumbo, 1999) for detecting differential item functioning (DIF) was the method used in this study to determine if ethnic-based differences existed in teacher ratings of students' performance on the language and literacy domain of a curriculum-embedded performance assessment. In this study, DIF would exist when students of equal ability who differ only by

ethnicity differ in how teachers rated them on a particular indicator. Although there are numerous procedures (e.g., Camilli & Congdon, 1999; Chang, Mazzeo, & Roussos, 1996; French & Miller, 1996; Hamilton, 1999; Miller & Spray, 1993; Potenza & Dorans, 1995; Welch & Hoover, 1993; Zwick, Thayer, & Mazzeo, 1997) available for the detection of DIF in polytomous items, the ordinal logistic regression procedure was selected because it included a method for determining the type of DIF, when DIF was detected, and it also included a method for computing effect sizes.

The ordinal logistic regression procedure consists of a test of statistical significance and an R^2 measure of effect size. For this study, a Type I error rate of .01 was set because of multiple DIF tests. The dependent variable was teacher ratings for each indicator. The independent variables were the total language and literacy domain score (i.e., the sum of the 12 language and literacy indicators for each student) and student ethnicity (coded: 1=African American, 2=White; where African American students was the focal group and White students was the reference group). Students' language and literacy domain scores ranged from 12 to 36.

A three-step modeling process was used to detect DIF. In the first model (null model), the total language and literacy domain score (the conditioning variable) was included to examine the extent to which the domain score influenced teacher ratings. In the second model, student ethnicity was entered into the model. Controlling for students' domain score, the inclusion of student ethnicity examined the extent to which student ethnicity influenced teacher ratings and to test for uniform DIF. An item shows uniform DIF if one of the groups, either African American students or White students, performed constantly at a higher rate, compared to the other group, across ability levels. In the third model (full model) the interaction term (i.e., the interaction between students' language and literacy domain score and students' ethnicity) was entered into the model. Therefore, in the full model, controlling for students' domain scores and students' ethnicity, the inclusion of the interaction term examined the extent to which the interaction between students' domain scores and students' ethnicity influenced teacher ratings and to test for non-uniform DIF. An item shows non-uniform DIF if the performance of one of the groups, either African American students or White students, depended upon ability level. Thus, there would be an interaction between the student ethnicity and ability level. To detect whether an indicator exhibited DIF, a two-degree-of-freedom Chi-squared test compared the full model to the null model. If the Chi-squared test had a p-value less than or equal to .01 and an R^2 difference of at least .13, the indicator exhibited DIF (Zumbo, 1999). Otherwise, there was not enough statistical evidence to indicate DIF for the indicator. Effect sizes were reported for statistically significant and non-significant DIF indicators. The Zumbo-Thomas measure of effect size for R^2 was computed by taking the difference between the R^2 values obtained from the full model and the null model (Zumbo, 1999).

Results

Teacher Rating of Students' Language and Literacy Performance

Table 1 shows the frequencies and percentages of teacher ratings of students' performance on the language and literacy domain. The internal consistency of teacher ratings on the domain was .96. Across items, the distribution of teacher ratings was negatively skewed. Thus, the majority of teacher ratings were in the *proficient* category. Teachers tended to rate students' performance highest on *Indicators* 4 and 5, reflecting speaking, and literature and reading, respectively. Across indicators, less than 9% of teacher ratings were in the *not yet* category.

Teacher Rating of Students' Performance by Ethnicity

Frequencies and percentages of teacher ratings of students' performance on the language and literacy domain by student ethnicity are represented in Table 2. Within student ethnicity, the percentage of students receiving ratings of *not yet* was greater for African American students than for White students. The percentage of White students receiving *not yet* ratings was between 2% and 6%, whereas between 4% and 10% of African American students received ratings of *not yet*. African American students receiving *not yet* ratings were approximately twice that of White students.

At the other extreme, the percentage of teacher ratings in the *proficient* category was greater for White students compared to African American students. The

Table 1
Frequencies (Percentages) of Teacher Ratings
of Students' Language and Literacy Performance (*n*=1,761)

	Rating		
Indicator	Not Yet	In Process	Proficient
1	97 (5.5)	728 (41.3)	936 (53.2)
2	149 (8.5)	708 (40.2)	904 (51.3)
3	94 (5.3)	671 (38.1)	996 (56.6)
4	76 (4.3)	636 (36.1)	1,049 (59.6)
5	87 (4.9)	579 (32.9)	1,095 (62.2)
6	146 (8.3)	685 (38.9)	930 (52.8)
7	134 (7.6)	699 (39.7)	928 (52.7)
8	154 (8.7)	690 (39.2)	917 (52.1)
9	119 (6.8)	792 (45.0)	850 (48.3)
10	140 (8.0)	778 (44.2)	843 (47.9)
11	93 (5.3)	760 (43.2)	908 (51.6)
12	113 (6.4)	720 (40.9)	928 (52.7)

Note. The internal consistency (Cronbach's alpha) of the scale was .96.

Table 2
Frequencies (%) of Teacher Ratings of Students' Language and Literacy Performance (*n*=1,761)

	Rating			
Indicator	Not Yet	In Process	Proficient	Total
1				
African American students	85 (5.9)	622 (43.1)	735 (51.0)	1,442
White students	12 (3.8)	106 (33.2)	201 (63.0)	319
Total	97 (5.5)	728 (41.3)	936 (53.2)	1,761
2				
African American students	130 (9.0)	612 (42.4)	700 (48.5)	1,442
White students	19 (6.0)	96 (30.1)	204 (63.9)	319
Total	149 (8.5)	708 (40.2)	904 (51.3)	1,761
3				
African American students	85 (5.9)	577 (40.0)	780 (54.1)	1,442
White students	9 (2.8)	94 (29.5)	216 (67.7)	319
Total	94 (5.3)	671 (38.1)	996 (56.6)	1,761
4				
African American students	69 (4.8)	543 (37.7)	830 (57.6)	1,442
White students	7 (2.2)	93 (29.2)	219 (68.7)	319
Total	76 (4.3)	636 (36.1)	1,049 (59.6)	1,761
5				
African American students	80 (5.5)	501 (34.7)	861 (59.7)	1,442
White students	7 (2.2)	78 (24.5)	234 (73.4)	319
Total	87 (4.9)	579 (32.9)	1,095 (62.2)	1,761
6				
African American students	129 (8.9)	594 (41.2)	719 (49.9)	1,442
White students	17 (5.3)	91 (28.5)	211 (66.1)	319
Total	146 (8.3)	685 (38.9)	930 (52.8)	1.761
7				
African American students	118 (8.2)	608 (42.2)	716 (49.7)	1,442
White students	16 (5.0)	91 (28.5)	212 (66.5)	319
Total	134 (7.6)	699 (39.7)	928 (52.7)	1,761
8				
African American students	137 (9.5)	598 (41.5)	707 (49.0)	1,442
White students	17 (5.3)	92 (28.8)	210 (65.8)	319
Total	154 (8.7)	690 (39.2)	917 (52.1)	1,761
9				
African American students	106 (7.4)	674 (46.7)	662 (45.9)	1,442
White students	13 (4.1)	118 (37.0)	188 (58.9)	319
Total	119 (6.8)	792 (45.0)	850 (48.3)	1,761
10				
African American students	125 (8.7)	659 (45.7)	658 (45.6)	1,442
White students	15 (4.7)	119 (37.3)	185 (58.0)	319
Total	140 (8.0)	778 (44.2)	843 (47.9)	1,761
11				
African American students	82 (5.7)	644 (44.7)	716 (49.7)	1,442
White students	11 (3.4)	116 (36.4)	192 (60.2)	319
Total	93 (5.3)	760 (43.2)	908 (51.6)	1,761
12				
African American students	98 (6.8)	612 (42.4)	732 (50.8)	1,442
White students	15 (4.7)	108 (33.9)	196 (61.4)	319
Total	113 (6.4)	720 (40.9)	928 (52.7)	1,761

percentage of White students receiving *proficient* ratings ranged from 58% to 74% whereas the percentage of African American students receiving *proficient* ratings ranged from 45% to 60%. Thus, White students' *proficient* ratings were approximately 1.25 times that of African American students.

Differential Item Functioning

The summary of the DIF analysis to detect item (indicator) bias is presented in Table 3. In the null model, the language and literacy domain score accounted for between 70% and 86% of the error variance in teacher ratings. Thus, only between 14% and 30% of the error variance in teacher ratings was unexplained. Controlling for the language and literacy domain score, the addition of student ethnicity into the model changed the R^2 between 0% and .16%. For instance, adding student ethnicity in the model for *Indicator* 9 (as reflected on the writing functional component) had no effect on the R^2. However, including student ethnicity in the models for the other indicators increased the R^2 between .01% and .16%.

Similarly, controlling for the language and literacy domain score and student ethnicity, the inclusion of the interaction term yielded differing effects. The amount of error variance controlled by the model by the inclusion of the interaction term ranged from 0% to .10%. For *Indicators* 5 and 10, adding the interaction term to the model reduced the R^2 by .04% and .01%, respectively. The indicators are in the areas of literature and reading, and writing. For *Indicators* 3, 4, 7, 8, 9, 11, and 12, the inclusion of the interaction term did not change the R^2. The indicators are reflected in all functional components except listening. For *Indicators* 1, 2, and 6, the inclusion of the interaction term increased the R^2 by .01%, .10%, and .02%, respectively. The increases occurred for indicators in the listening, and literature and reading functional components.

To test for statistical significance, a two-degree of freedom DIF Chi-squared test was computed for each indicator. As indicated by the results in the table, the amount of variation in teachers' ratings explained by the inclusion of students' domain scores, ethnicity, and the interaction term was either nil or small. Therefore, there was not enough evidence to detect a significant ethnic-based difference in teacher ratings for each indicator ($p > .01$). The magnitude for the non-significant differences in the ratings ranged from 0 to .0016. Thus, the magnitude of the observed ethnic-based differences in teacher ratings is less than the .13 suggested by Zumbo (1999).

Discussion and Conclusion

This study examined the extent to which ethnic-based differences exist in teacher ratings of African American students and White students on the language and literacy domain of a curriculum-embedded performance assessment for students in grade 1. It extended previous research on performance assessments to focus on issues related to equity in teacher ratings on a curriculum-embedded performance assessment of young children in an urban school district. The major findings of this study are discussed in terms of the: (a) consistency in teacher ratings, (b) dispar-

Table 3
Summary of DIF Analysis for Predicting Teacher Ratings on the Language and Literacy Domain (n=1,761)

	R^2 for Predictors				
Indicator	LL Score	LL Score + Ethnicity	LL Score + Ethnicity+ LL Score* Ethnicity	DIF c^2(2) Test	ES
1	.7195	.7197	.7198	.884*	.0003
2	.7144	.7147	.7157	2.558*	.0013
3	.7026	.7028	.7028	.466*	.0002
4	.7191	.7193	.7193	.340*	.0002
5	.7559	.7564	.7560	1.895*	.0001
6	.8160	.8165	.8167	1.799*	.0007
7	.8534	.8539	.8539	1.747*	.0005
8	.8375	.8384	.8384	2.797*	.0009
9	.8379	.8379	.8379	.305*	.0000
10	.8506	.8507	.8506	.450*	.0000
11	.7775	.7786	.7786	3.369*	.0011
12	.7935	.7951	.7951	4.781*	.0016

Note. LL = Language and Literacy
*p > .01.

ity in teacher ratings of African American and White students, and (c) absence of indicator bias on the language and literacy domain.

The first major finding of the study indicated that across indicators on the language and literacy domain, teachers demonstrated a high level of consistency in rating students' performance. The internal consistency index (Cronbach's alpha) was .96. Thus, only about 4% of the observed ratings could be attributed to systematic error. The reliability index in this study is consistent with the findings of Meisels et al. (1995) in which the reported spring reliability index for 17 items on the language and literacy domain of the Work Sampling System was .94. The high reliability of teacher ratings could be attributed to teacher training on the use of the instrument, the accessibility of developmental guidelines for indicators on the checklist, teacher familiarity with the curriculum, and ongoing observations. Prior to the 2001-2002 administration of the assessment, teachers received training on the use of its developmental guidelines and checklists. As stated earlier, the developmental guidelines and checklists provided a set of criteria for teachers to use as they rated students' performance. The use of scoring rubrics that specifically identifies the criteria for scoring is aligned with the *Standards* (AERA, APA, & NCME, 1999). Furthermore, teachers had multiple opportunities, over time, to observe students' performance on the indicators prior to the spring 2002 data collection period. The cognitive development of children in the early grades varies. There are continual

changes in children's acquisition of reading and language skills that are not accurately captured with traditional multiple-choice tests. Multiple opportunities for teachers to observe students demonstrating language and literacy skills provide a holistic understanding of the progression of students' learning and the interaction of reading and language skills. Through continual observations, teachers were in a better position to draw informed conclusions about the level at which students were able to demonstrate the competency for each indicator.

The second major study finding was the discrepancy in teacher ratings between African American and White students. Across indicators, a greater proportion of African American students received ratings in the *not yet* category compared to White students. Furthermore, a smaller proportion of African American students received ratings in the *proficient* category compared to White students. At first glance, the results are disturbing. However, the results are consistent with findings related to achievement gaps between African American and White students on standardized tests. For example, Lee (2002) conducted a comprehensive review of available National Assessment of Educational Progress (NAEP) data to explore Black-White and Hispanic-White achievement gaps in reading and mathematics over a 30-year period. Lee found that achievement gaps narrowed in the 1970s and 1980s but then stabilized or widened in the 1990s. Several factors may affect ethnic achievement gap trends. These include changes in socioeconomic and family conditions, youth culture and student behavior, and schooling conditions and practices (Lee, 2002).

The final major finding of this study was the absence of DIF in teacher ratings on the indicators. The results showed that the language and literacy domain score was the major contributor to teachers' predicted ratings, accounting for between 70% and 86% of the total variation in teacher ratings. The inclusion of student ethnicity and the interaction term in the models explained only a small amount of the total amount of variation in teachers' ratings. Thus, there was not enough statistical evidence to support ethnic-based differences in teacher ratings of students on the language and literacy domain. The results of this DIF study are consistent with technical documentation of the absence of ethnic-based DIF on the spring 2001 statewide administration of the curriculum-embedded assessment for students in grade 1. Considering that this study was conducted in an urban school district where the percentage of teachers was 52% White and 47% African American and the percentage of African American students was 82%, teacher sensitivity to the student demographic make up of the district may have reduced the potential for teachers to demonstrate partiality toward a particular group. The findings further suggest that teacher biases do not always play out in assessments—which are contradictory to other assertions (Ferguson, 1998, Ford, 1996; Ford et al., 1999).

The findings in this study are to be viewed with caution in terms of generalizability. Again, this study explored ethnic-based differences in a particular state-wide assessment program in a specific urban school district. The curriculum standards exhibited in this state's assessment program are not necessarily reflected in another

state's assessment program. Furthermore, this study's focus was a single school district's performance on a state-wide assessment program. The lack of representation of first-grade students from across the state makes the findings of this study not representative of the diverse geographical and economic make up of the state. Thus, the findings in this study cannot be generalized to other state assessment programs and are restricted to the urban school district used in this study.

Implications

The use of educational assessments in American education has a long and rich history. The current educational environment with its focus on improving achievement offers yet a great opportunity to transform public education for African American students and other educational vulnerable student populations. But, school leadership must play a more significant role in the educational transformation efforts to improve school outcomes for African American students and reduce the longstanding Black-White achievement gaps found in many of the nation's public schools.

To ensure that teachers' biases do not negatively affect students' academic work, including assessments, it is important that school leadership offers ongoing teacher professional development to assist teachers with implementing assessments, as well as evidence-based pedagogy and curricula in the classroom. Toward this end, it is essential that teachers learn how to develop classroom behaviors that do not show favoritism but allow them to fairly and consistently interact with all students. This type of professional development helps teachers become aware of and change behaviors that often affect students' school efficacy and motivation to "work hard" to achieve. It is also essential that administrators and other school leaders stress to teachers the importance utilizing the information gained from professional development. Additionally, it is imperative that teachers are given ongoing support and feedback on their progress, as they relate to students' outcomes in their classes.

References

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

Baker, E. L., & O'Neil, H. F., Jr. (1995). Diversity, assessment, and equity in educational reform. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 69-88). Boston: Kluwer Academic Publishers.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*, 323-341.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Chubb, J. E., & Loveless, T. (Eds.). (2002). *Bridging the achievement gap*. Washington, DC: Brookings Institution Press.

Cobia, D. C., & Henderson, D. A. (2007). *Developing an effective and accountable school*

counseling program (2nd ed.). Upper Saddle, NJ: Pearson.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston: Kluwer Academic Publishers.

Dichtelmiller, M. L., Jablon, J. R., Meisels, S. J., Marsden, D. B, & Dorfman, A. B. (1998). *Using Work Sampling guidelines and checklists: An observational assessment*. Ann Arbor, MI: Rebus.

Ferguson, R. F. (1998). Teachers' perceptions and expectations and the black-white test score gap. In C. Jencks & M. Phillips, (Eds.), *The black-white test score gap* (pp. 318-374). Washington, DC: Brookings Institution Press.

Flowers, L. A., Milner, H. R., & Moore, J. L., III. (2003). Effects of locus of control on African American high school seniors' educational aspirations: Implications for Preservice and inservice high school teachers and counselors. *The High School Journal, 87,* 39-50.

Ford, D. Y. (1996). *Reversing underachievement among gifted black students: Promising practices and programs.* New York: Teachers College Press.

Ford, D. Y., Grantham, T. C., & Bailey, D. Y. (1999). Identifying giftedness among African American males: Recommendations for effective recruitment and retention. In V. C. Polite & J. E. Davis (Eds.), *African American males in school and society: Practices & policies for effective education* (pp. 51-67). New York: Teachers College Press.

Ford, D. Y., & Moore, J. L., III. (2004). The achievement gap and gifted students of color. *Understanding Our Gifted, 16,* 3-7.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Friedman, T. L. (2005). *The world is flat*. New York: Farrar, Straus & Giroux.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*, 211-235.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Washington, DC: Brookings Institution Press.

Kozol, J. (1991). *Savage inequalities: Children in America's schools.* New York: Harper Perennial.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, *31*(1), 3-12.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

Meisels, S. J. (1993). Remaking classroom assessment with the Work Sampling System. *Young Children, 48*(5), 34-40.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*, 73-95.

Meisels, S. J., Dorfman, A., & Steele, D. (1995). Equity and excellence in group-administered and performance-based assessments. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 221-242). Boston: Kluwer Academic Publishers.

Meisels, S. J., Liaw, F., Dorfman, A., & Nelson, R. F. (1995). The Work Sampling System: Reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly, 10*, 277-296.

Meisels, S. J., Jablon, J. R., Marsden, D. B., Dichtelmiller, M. L., & Dorfman, A. (1994). *The Work Sampling System* (3rd ed.). Ann Arbor, MI: Rebus.

Moore, J. L., III. (2003). Introduction. *The High School Journal, 87,* 1-3.

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Governors' Association. (1986). *Time for results: The governors' 1991 report on education*. Washington, DC: Author.

Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?* (NCES Working Paper Series No. 96-08). Washington, DC: U.S. Department of Education. Available online at http://nces.ed.gov/pubs96/9608.pdf

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment in polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Sears, S. J. (2002). This issue. *Theory Into Practice, 41,* 146-147.

Southern Education Foundation. (1995). *Redeeming the American promise: Executive summary.* Atlanta, GA: Author.

Southern Education Foundation. (1999). *Miles to go Maryland.* Atlanta, GA: Author.

Southern Education Foundation. (2000a). *Miles to go Arkansas.* Atlanta, GA: Author.

Southern Education Foundation. (2000b). *Miles to go South Carolina.* Atlanta, GA: Author.

Usdan, M. (2006). Mayors and public education: The case for greater involvement. *Harvard Educational Review, 76,* 147-151.

U. S. Department of Education. (2001). *No Child Left Behind Act of 2001 (H. R. I.).* Washington, DC: Author.

U.S. Department of Education. (2004). *A guide to education and No Child Left Behind.* Washington, DC: U.S. Government Printing Office.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1-19.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available online at http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

Zwick, R., Thayer, D. T., Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-344.