Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis*, *19*(2), 83-97. © 1997 Sage Publications, Inc.

Gender and Racial/Ethnic Differences on Performance Assessments in Science

Stephen P. Klein, Jasna Jovanovic, Brian M. Stecher, Dan McCafferey, Richard J. Shavelson, Edward Haertel, Guillermo Solano-Flores, & Kathy Comfort

We examined whether the differences in mean scores among gender and racial/ethnic groups on science performance assessments are comparable to the differences that are typically found among these groups of traditional multiple-choice tests. To do this, several hands-on science performance assessments and other measures were administered to over 2,000 students in grades five, six, and nine as apart of a field test of California's statewide testing program. Girls tended to have higher overall mean scores than boys on the performance measures, but boys tended to score higher than girls on certain types of questions within a performance task. In contrast, differences in mean scores among racial/ethnic groups on one type of test (or question) were comparable to the differences among these groups on the other measures studied. Overall, the results suggest that the type of science test used is unlikely to have much effect on gender or racial/ethnic differences in scores.

Proponents of education reform often recommend replacing traditional multiple-choice tests with performance assessments (National Commission on Testing and Public Policy, 1990). A factor cited in support of this recommendation is the consequences for curriculum and instruction that may stem from relying solely on multiple-choice tests. Specifically, education reformers contend that such tests emphasize factual knowledge to the exclusion of important process and application skills (Frederiksen, 1984; Glaser, 1988).

This is an especially serious concern among those who are working to improve science education (Wiggins, 1989). They argue that if instruction continues to be focused on raising scores on traditional achievement tests, students will not learn the skills that are integral to a conceptual understanding of science (Shavelson, Carey, & Webb, 1990). However, if a wider range of curricular relevant abilities is measured in large-scale, high-stakes testing programs, then teachers will be motivated to emphasize these skills in their classrooms (Shavelson, Baxter, & Pine, 1992; Tamir, 1993). As outlined in a policy report issued by the National Center for Improving Science Education (Raizen et al., 1989), the goal of standards-based alternative assessments in science is to reinforce or even drive curricular reforms. To do this, the assessments must involve students interacting with materials and engaging in activities that stimulate problem-solving and critical thinking while still promoting the acquisition of factual knowledge.

Another concern is that females and students of color usually score much lower than Whites on traditional multiple-choice tests. For example, on the science section of the 1990 National Assessment of Educational Progress (NAEP), the mean score for 4th-grade boys was two points (2.0 standard errors) higher than the mean score for girls (Jones, Mullis, Raizen, Weiss, & Weston, 1992). This gender difference widened significantly in grades 8 and 12. Mean score differences are even larger among racial/ethnic groups. For example, the mean for 4th-grade Whites on the NAEP science

test was 30 points higher than the mean for Blacks. (Standard errors for Whites, Hispanics, and Blacks were 1.0, 1.5, and 1.5, respectively).

Several hypotheses have been advanced to explain gender differences in test scores. One theory is that the multiple-choice format itself favors males (Bolger & Kellaghan, 1990; Harding, 1981; Murphy, 1982). That is, multiple-choice tests reward students who guess (Rowley, 1974; Slakter, 1968), and boys are more willing to take this risk (Ben-Shakkar & Sinai, 1991; Hanna, 1986). Females on the other hand are more likely to choose an "I don't know" response option, particularly on physical science questions (Linn, Benedictis, Delucchi, Harris, & Stage, 1987). However, girls score significantly higher than boys on multiple-choice reading tests (Langer, Applebee, Mullis, & Foertsch, 1990; Pinell et al., 1995), so it is not clear that format alone is the answer.

It also has been postulated that students do better on test items that deal with objects or events that are drawn from their own "sphere of experience" (Erickson & Erickson, 1984). If so, boys would have an advantage on tasks that are sensitive to experience with science-related activities, such as tinkering with mechanical objects and participating in science clubs (S. Johnson, 1987; Jones et al., 1992). At the same time, gender-related differences in specific cognitive abilities (e.g., spatial ability) may lead boys and girls to perform differently on certain test items (Linn et al., 1987).

There are several explanations for the observed differences in mean test scores among racial/ethnic groups. Some of these differences appear to be related to home and school characteristics. For example, high mathematics and science scores are associated with having more family resources and learning opportunities in the home. This relationship holds for all groups, but Whites tend to have more of these advantages than Blacks or Hispanics (Peng, Wright, & Hill, 1995). In addition, many minority students are more likely than Whites to attend schools with the following characteristics that are associated with lower performance: poor school climate, less-qualified teachers, low curriculum requirements, less press for achievement, and more "low-track" programs. Minority students also are generally less likely to be ready for school, have lower academic expectations, are less engaged in learning, and take fewer advanced courses. When taken together, these home, school, and individual factors are associated with about 45% of the variation among groups in mean NAEP mathematics and science scores (Peng et al., 1995). The remaining variation is less understood.

Several researchers have explored whether differences in mean scores among racial/ethnic groups stem from certain questions with a test that are especially troublesome for minority students. These "differential item functioning" studies generally find that if a question is relatively hard or easy for one group, it has that same characteristic in all of the other groups studied. However, researchers are usually at a loss to explain why the few aberrant questions behave as they do (Zwick & Ercikan, 1989). Whatever the reason for the large differences among racial/ethnic groups, they do not appear to stem from readily observable item characteristics.

It has been suggested that performance assessments will reduce differences among groups by reinforcing appropriate curriculum changes and by providing students with hands-on opportunities to demonstrate their knowledge and understanding of scientific principles, not simply by recalling facts, but by constructing solutions (Shavelson, 1997). These measures emphasize the process by which students generate

solutions, not just the correctness of the solution itself (Baxter, Shavelson, Goldman, & Pine, 1992; Carey & Shavelson, 1989). The underlying theory is that individuals approach problem-solving differently because of varying styles, not differing abilities (Paris, Lawton, Turner, & Roth, 1991). Accordingly, proponents of performance assessments expect that these measures will narrow the differences in scores among groups because they are designed to allow for this individual variation (Neil & Medina, 1989) and they put less emphasis on guessing, exposure to science-related activities outside the classroom, testwiseness, and other presumably extraneous factors (Jenkins & MacDonald, 1989; M. Johnson, 1990).

Whether performance assessments will in fact reduce gender and racial/ethnic differences in science test scores is an open question. Jovanovic and her colleagues (Jovanovic & Shavelson, 1995; Jovanovic, Solano-Flores, & Shavelson, 1994) explored science achievement in several content domains (e.g., physics, chemistry, earth science) with performance assessments and traditional testing methods (i.e., multiple choice, short answer). They found that males and females generally had similar means on both types of measures. The few significant differences that emerged depended on the specific science content domain assessed. For example, girls had a slight advantage on tasks related to earth science and ecology (e.g., classification of leaves and rocks), whereas boys had an advantage on activities related to electricity (Jovanovic & Shavleson, 1995). These differences occurred regardless of the method of measurement used. Hence, students' prior science-related experiences may play a role regardless of the type of test used (Jovanovic et al., 1994).

Relatively little is known about whether performance assessments will reduce the racial/ethnic differences that are found with multiple-choice tests. The few studies that have been done suggest that they will not have much effect (Linn, Baker, & Dunbar, 1991). For example, the 1992 NAEP mathematics assessment contained both regular (short) constructed response tasks and extended response items (Peng et al., 1995). There were considerable differences in scores in favor of White students over Hispanic or Black students. Among eighth-grade students, Whites were two to three times more likely than Hispanics or Blacks to correctly answer the two regular constructed response tasks. The difference were even more dramatic on the extended-response task, where 49% of White eight-grade students gave at least a minimal response compared to 16% of Hispanic students and 13% of Black students (Mullis, 1994). Similarly, in 1992, NAEP conducted a supplemental assessment of fourth-grade students' oral reading proficiency (including accuracy, rate, and overall literacy development) in addition to the regular NAEP assessment of reading. This study found that the gap between White and minority students on traditional measures of reading corresponded to the gap between them in oral reading (Pinnell et al., 1995). To our knowledge, there are no published studies that compare the scores of racial/ethnic groups on performance assessments and multiple-choice tests in science.

Purpose of the Study

We investigated whether the use of hands-on science performance assessments, such as in large-scale testing programs, is likely to affect the differences in scores among gender and racial/ethnic groups that are typically found on traditional standardized multiple-choice tests. To do this, we examined students' science achievement when

measured by both performance assessments and traditional multiple-choice tests. We also examined whether certain types of performance tasks (and question types within tasks) are more likely than others to affect these disparities and whether these results are consistent across grade levels.

Method

Participants

Several hands-on performance assessments and other measures were administered to students in grades five, six, and nine in conjunction with a 1993 field test of the California Learning Assessment System (CLAS). All together, the field test involved over 2,400 students from 90 classrooms across 30 schools. Table 1 presents a breakdown of the sample by grade level, gender, and racial/ethnic group. With minor exceptions because of absences, all students completed a five-classroom-period test battery.

TABLE 1
Sample Characteristics by Grade Level

Grade level	Number of		Mean *N* per measure	Percent female	Percent Black	Percent Hispanic	Percent Asian
	Schools	Classrooms					
Five	16	38	1,003	52%	4%	21%	14%
Six	16	38	1,018	51%	11%	21%	21%
Nine	3	16	382	54%	12%	22%	20%

Note: The total sample of participating students was as follows: grade five = 1,089, grade six = 1,121, and grade nine = 420.

A different battery was used at each grade level (Table 2). In addition, all students were evaluated by their teacher for "overall ability in science" relative to the other students in their classroom. A five-point scale was used for this purpose. The fifth- and sixth-grade test batteries also included the 35 minute multiple choice science subtest of the Iowa Tests of Basic Skills (ITBS) that was appropriate for these grade levels (Hoover, Hieronymus, Frisbie, & Dunbar, 1994).

TABLE 2
Measures Used at Each Grade Level

Measures (and tasks within investigations)	Grade five	Grade six	Grade nine
Teacher ratings	X	X	X
ITBS science	X	X	
CLAS other			
Multiple choice	X	X	X
Justified multiple choice	X	X	
Open-ended	X	X	
CLAS hands-on investigations			
Investigation #1 (rocks, roads, & critters)	X	X	
Investigation #2 (fish, cooling, & erosion)			X
Shell-based investigations			
Incline (plan-perform & analyze-apply)	X		
Friction (plan-perform & analyze-apply)	X		
Classification (animals & materials)		X	
Inference (levers & pendulums)		X	
Radiation (design & analysis)			X
Rate of cooling (design & analysis)			X

Note: Each shell-based investigation had two tasks. The grade five shell-based tasks were developed by a team of researchers at the University of California, Santa Barbara, the grade six tasks by a RAND team, and the grade nine tasks by a Stanford University/Far West Regional Laboratory team.

Measures

One challenge to developing performance assessments for large-scale use is the need to produce multiple versions of a test to assess a particular science knowledge domain (such as force and motion). This is an especially important consideration when the tests are highly "memorable." We addressed this problem by constructing multiple measures from generalized task development "shells." Originally created for the development of multiple-choice items, shells have been defined as "hollow" questions whose syntactic structure allows for the generation of similar questions (Haladyna & Shindoll, 1989). This idea has been extended to include any template for constructing questions that is based on structural specifications of a knowledge domain (Bormuth, 1971; Hively, Patterson, & Page, 1968).

To facilitate the discussion that follows, a "shell" is defined as a set of specifications for generating one or more tasks. A shell describes the critical feature of an assessment, such as its structure, the types of variables involved, and the cognitive and procedural demands placed on students (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1997). A "task" consists of a set of activities in which the student engages and a set of questions that the student answers (in writing) about these activities.

Tasks developed from the same shell have many important characteristics. In principle, this allows for the comparison of student performance on conceptually parallel measures (Klein et al., 1996). An "investigation" consists of one or more tasks that deal with a given topic. For example, one of the grade five investigations was called "incline" because both of its tasks dealt with incline planes. The first of these two tasks, *"plan-perform,"* was derived from a shell that involved "planning, designing, and performing" activities. The second task in this same investigation, *"analyze-apply,"* was developed from an "analyzing, interpreting, and applying" shell. These same two shells also were used to create two different but seemingly parallel tasks for another grade five investigation that dealt with the topic of friction. The measures used at each grade level are listed in Table 2 and described briefly below. (See Stecher & Klein, 1995, for more complete descriptions of all the measures used in this research.)

Fifth-Grade Test Battery

The fifth-graders completed two shell-based investigations, incline and friction. Both investigations required one classroom period, had two tasks apiece (plan-perform and analyze-apply), and dealt with the general science topic of force and motion. In the incline investigation, the student examined the relationship between the steepness of a ramp and the amount of force needed to move a toy truck up it. IN the friction investigation, the student examined the relationship between the roughness of various surfaces and the amount of force needed to pull different-sized blocks of wood across each surface.

A third classroom period was devoted to three hands-on tasks developed by CLAS. In Task 1 of the CLAS hands-on investigation, the student used various tools to determine which of three types of *rocks* would work best for the surface of outdoor picnic tables and benches. Task 2 had the student conduct an experiment to determine whether it would be better to use a paved or gravel *road* to remove debris from a work site. In Task

3, students were given a bag of "animals without backbones." They were then asked to create their own system for classifying these *"critters"* and explain why each one belonged in the category to which the student placed it. Within a classroom, students were assigned randomly to tasks. After time was called for one task, they switched work stations and began the next task so that by the end of the period, all students had worked on all tasks. The fifth-graders also took a CLAS test that had 24 multiple-choice items, 3 justified multiple-choice items (i.e., questions in which the student not only selected a choice but also explained the rationale for this selection), and 2 "open-ended" questions.

Testing was conducted over a five-day period. Students took two shell-based tasks from one investigation on Day 1 and two shell-based tasks from another investigation on Day 3. Half of the classrooms took the incline investigation first and the other half took friction first. The CLAS hands-on task was administered on Day 2, the ITBS science test on Day 4, and the other CLAS measures on Day 5. Teachers completed their ratings without knowing how well their students did on any of the measures.

Sixth-Grade Test Battery

The sixth-grade test battery contained two hands-on tasks generated from a "classification" shell and two hands-on tasks from an "inference" shell (however, unlike the fifth- and ninth-grade test plans, the sixth-grade shell-based tasks were not nested within investigations). Each sixth-grade shell-based task took one half of a classroom period. The classification shell involved a brief tutorial (called the "tuning" activity) on how to construct a 2 x 2 classification system. Students were then asked to construct their own 2 x 2 system for classifying eight objects. Task 1 involved classifying *animals* (including fur, bone, shell, and rock).

The inference shell required students to gather and record data about two independent variables, use these data to make an inference about which variable had the greater effect on a dependent variable, and then predict the value on the dependent variable for a combination of the two independent variables they could observe but not test. In Task 1, the student conducted an experiment to determine whether the length of a *pendulum* or its mass had a greater effect on the pendulum's period. In Task 2, the student conducted an experiment to determine whether the length of a *lever* or the relative position of its fulcrum point had a greater effect on the force needed to lift an object. They also estimated the force needed to lift this object with a lever they could observe but not test. A student took one inference and one classification task in one classroom period and the other inference and classification task in another period.

The sixth-graders took one classification task and one inference task on Day 1, the other pair of shell-based tasks on Day 3, the three CLAS hands-on tasks described above on Day 2, the ITBS science test on Day 4, and the CLAS multiple-choice, justified multiple-choice, and open-ended measures on Day 5. A counterbalanced design was used for the shell-based tasks so that about one fourth of the classrooms were assigned to each of the four possible sequence combinations (e.g., animals and levers on Day 1 and materials and pendulums on Day 3). On Days 1 and 3, about half the students in a classroom began with a classification task while the other half worked on an inference task. They then switched tasks halfway through the period.

Ninth-Grade Test Battery

The ninth-grade battery included two shell-based investigations, radiation and the rate of cooling, both of which dealt with the transfer of heat and energy. Each investigation included two tasks (*design* and *analysis*) and each task took one classroom period. In one radiation task, students designed an experiment to test the relationship between colors and rate of heat absorption. This included identifying the factors (such as water volume, distance from heat source, etc.) to be controlled, describing the methods used, and designing a chart on which to record results. In a subsequent radiation task, students analyzed the results of an experiment. This involved using an equation that related heat, temperature change, and volume to infer a solution to the practical problem that motivated the experiment.

In the rate-of-cooling investigation, one task involved designing an experiment to test the effects of different fabrics on heat loss. In a subsequent task, students analyzed the results of an experiment (including the use of an equation that relates to the relationships among heat, temperature change, and volume).

The 9th-graders also took a CLAS-developed, 21-item, multiple-choice science test and performed a three-task, one-classroom-period, hands-on investigation. Task 1 of this investigation examined the reason *fish* were dying in a lake, Task 2 involved a *rate-of-cooling* experiment, and Task 3 was a rock *erosion* problem. Although the CLAS tasks were designed for a statewide assessment of 10th-graders, they were judged to be appropriate for students in grades 9-12. About half of the 9th-grade classrooms took the two shell-based radiation tasks, the CLAS measures, and then the two shell-based, rate-of-cooling tasks while the other half of the classrooms took these measures in the opposite sequence.

Procedures

The testing sessions were conducted by specially trained “exercise administrators” (although the classroom teacher remained in the room). Portable partitions were used with the hands-on tasks so students could not observe or interact with one another while completing the tasks. Testing typically occurred in the cafeteria or other large room at the school so that each student had enough space to work with a task’s equipment and materials. Some hands-on tasks required an entire classroom period, while others were allocated one half or one third of a period. A few of the shorter tasks were administered simultaneously within a classroom (i.e., some students took one task while others took a different task until time was called, at which point students rotated to a different work station for the next task).

Students recorded their responses to the open-ended measures in test booklets as they worked their way through tasks. This booklet contained directions, a list of the materials and equipment the student needed for the task, and several separately scored questions. There was considerable variation in the nature of these questions, even within the same task. A few could be answered with a single word or number, but most required filling in a table, constructing a figure, providing a short written explanation of the results obtained, or using the results to draw conclusions and make inferences about the solution

to a related problem. Some required writing one or more short paragraphs. (See Stecher & Klein, 1995, for copies of all the measures and scoring guides used in this research.)

At all three grade levels, the sequence in which the measures were administered was counterbalanced across schools. All the tasks within an investigation were taken before the students took another measure, and there was always at least one non-shell-based measure administered between investigations.

Scoring

There was one team of readers for each hands-on performance task. Team size varied from 5 to 16 readers as a function of number of students taking a task and the time it took to grade the responses to it. Almost all the readers were teachers. There was a separate, semi-analytic scoring rubric for each task. A rubric usually consisted of a set of rules for assigning points to possible types of student responses to indicated how close each type came to the model answer. Readers were trained and supervised by project staff. A stratified random sampling plan was used to assign student answer booklets to "batches" so that no batch contained more than one answer from a given classroom. Batches were then assigned randomly to readers. Readers were not informed of student characteristics.

The total raw score on a task was the sum of the points the students received on that task. We did not assign weights to questions within tasks, so questions were effectively waited by their standard deviations. The total scores on a task were converted to z-scores for the analysis discussed below.

Results

The scoring was very reliable. There were no significant differences in means between readers and the median correlation between two independent readers on a shell-based task was 0.95. The ITBS score distributions in our samples of fifth- and sixth-graders coincided very closely with the national norm group. For example, the 25^{th}, 50^{th}, and 75^{th} percentile points in our sixth-grade sample corresponded to the 25^{th}, 50^{th}, and 73^{rd} percentile points nationally. The correlation between tow classroom periods of performance assessments (such as between the grade five friction and incline investigations) ranged from 0.62 to 0.80. (The median was 0.75.) Tables 3 and 4 contain inter-reader correlations and score reliabilities for each measure, respectively. To facilitate comparisons between groups on different but similarly reliable measures, we constructed a total performance assessment score (PA) for each student. This score was the sum of the student's scores over three classroom periods of performance assessments. For example, a fifth-grader's total PA score was the sum of that student's scores on friction, incline, and CLAS hands-on investigations (with each task being weighted equally in computing this total).

TABLE 3
Inter-Reader Correlations by Task and Grade Level

Measure	Grade five	Grade six	Grade nine
Incline (tasks 1 & 2)	0.95	—	—
Friction (tasks 1 & 2)	0.94	—	—
CLAS hands-on—rocks	0.95	0.88	—
CLAS hands-on—roads	0.86	0.83	—
CLAS hands-on—critters	0.85	0.85	—
CLAS open-ended	0.79	0.87	—
CLAS justified multiple choice	0.77	0.81	—
Inference—pendulums	—	0.95	—
Inference—levers	—	0.94	—
Classification—tuning module	—	0.97	—
Classification—animals	—	0.93	—
Classification—materials	—	0.92	—
Design—radiation	—	—	0.94
Design—rate of cooling	—	—	0.96
Analysis—radiation	—	—	0.96
Analysis—rate of cooling	—	—	0.97
CLAS hands-on—Lake Wilmar	—	—	0.87

Note: In grade six, the same tuning module was used for both classification tasks. The tabled value corresponds to the first time the student took this module.

TABLE 4
Score Reliability for One Classroom Period

Measure	Grade five	Grade six	Grade nine
Internal consistency reliability			
ITBS—science	0.82	0.84	—
CLAS other	0.69	0.65	0.61
CLAS hands-on	0.78	0.75	0.44
Alternate form reliability			
Friction & incline	0.75	—	—
Inference + classification	—	0.69	—
Radiation & rate of cooling (design)	—	—	0.62
Radiation & rate of cooling (analysis)	—	—	0.80

Notes: The coefficient alphas for the CLAS multiple choice, justified multiple choice, and open-ended sections of CLAS other were 0.76, 0.59, and 0.49 respectively at grade five and 0.76, 0.59, and 0.57 at grade six. The reliability of the "inference + classification" score was the mean correlation between all of the possible pairs of these tasks when the two tasks within a pair came from different shells (e.g., correlation of pendulums + animals with levers + materials). At grade nine, there was a 0.84 correlation between radiation and rate of cooling total scores (design + analysis), but each of these total scores was based on two classroom periods of testing.

Gender Differences

Figure 1 uses "box-and-whisker" plots to display the relationships between test type and gender. There are five vertical lines on each plot. Reading from left to right, these lines correspond to the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentile points in each group. In grade six, girls had slightly higher performance assessment (PA) scores than boys even though girls and boys had nearly identical distributions of ITBS science scores.

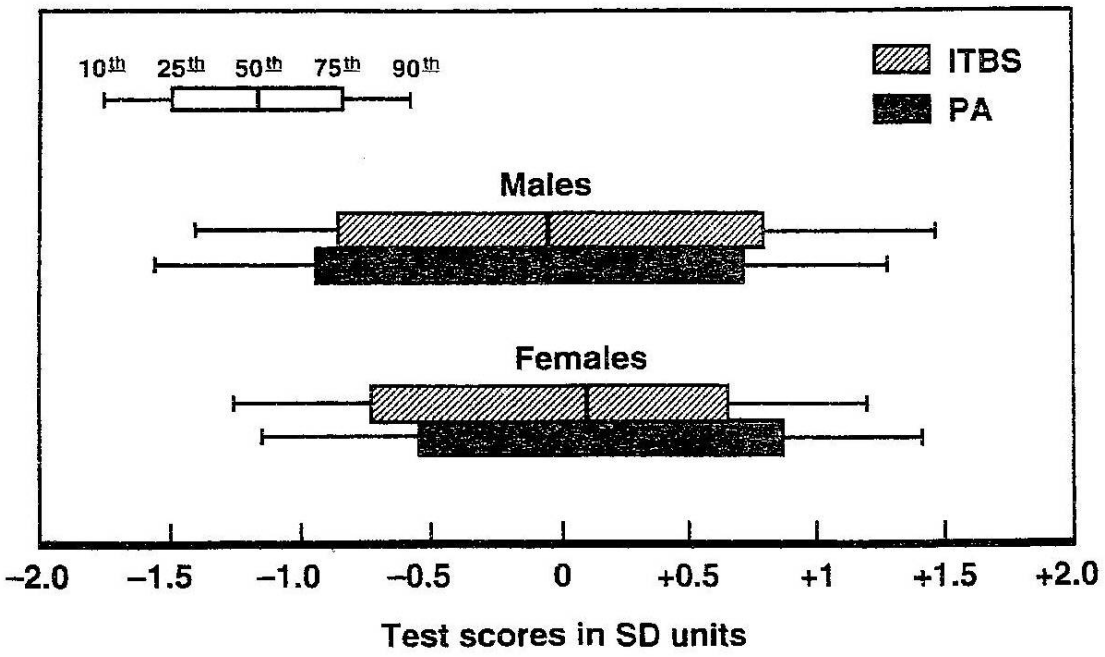

FIGURE 1. *Distribution of scores on performance assessments (PAs) and the ITBS science test by gender for sixth-graders.*

There were statistically significant differences in mean scores between gender groups on all the PAs in grade six but not in grade five or nine (Table 5). Girls also had consistently higher mean teacher ratings than boys. The only time boys had a significantly higher mean than girls was on the grade-nine CLAS multiple-choice test. Girls and boys had nearly identical correlations among the measures at each grade level. (Corresponding coefficients were usually within 0.05 of each other.)

TABLE 5
Girls' Mean Minus Boys' Mean in Z-Score Units

Measure	Grade five	Grade six	Grade nine
Teacher rating	0.15*	0.33*	0.27
ITBS science	–0.03	0.02	—
CLAS other	–0.07	0.05	–0.30*
CLAS hands-on	0.28*	0.42*	0.08
Incline	0.07	—	—
Friction	0.11	—	—
Classification	—	0.25*	—
Inference	—	0.17*	—
Radiation	—	—	0.06
Rate of Cooling	—	—	0.02

*Difference between girls and boys significant at $p < .05$. The CLAS hands-on score was based on three tasks. The score on each of the other six hands-on investigations was based on two tasks apiece.

Racial/Ethnic Differences

Figure 2 shows that roughly 75% of the Hispanic and Black fifth- and sixth-graders are in the bottom one third of the distribution of Whites. (We combined data across these two grade levels to provide reasonably stable estimates of the five percentile points shown on each plot.) Only about 10% of the Black and Hispanic students had ITBS or PA scores that exceeded the White median score. Figure 2 also shows that the differences among racial/ethnic groups on the ITBS science subtest are almost identical to the differences among them on the performance measures. In short, the use of performance assessments did not narrow or widen the gap in scores between groups.

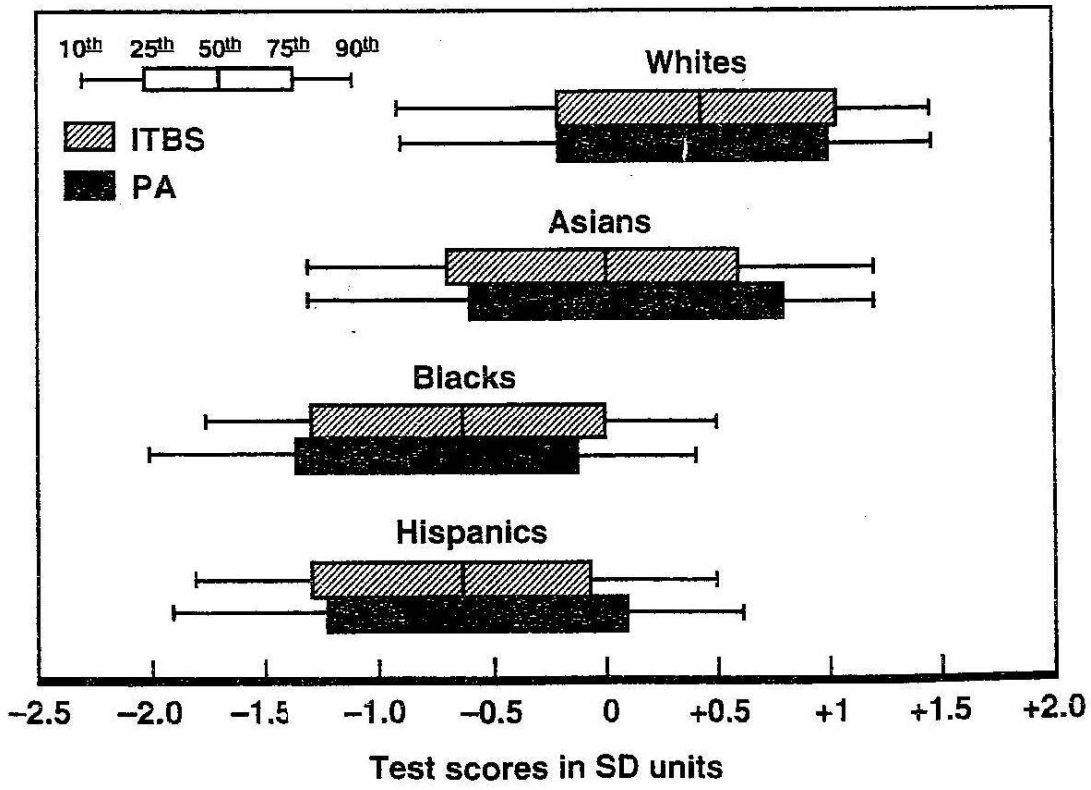

FIGURE 2. *Distribution of scores on performance assessments (PAs) and the ITBS science test by racial/ethnic group for fifth- and sixth-graders.*

The differences in means (in standard score units) between racial/ethnic groups for one classroom period of testing with each type of measure are presented in Table 6. These data indicate that Black and Hispanic students had means that were substantially lower than those of Whites. Grade five and six Asian students had slightly lower mean test scores than Whites (but only in grade six were the differences significant). Interesting, the grade-nine Asian students had a significantly higher mean teacher rating than their White classmates.

Some schools had substantially more minority students than others. Thus, differences among groups may be due to differences in school quality that are related to race and/or ethnicity rather than racial/ethnic group per se. When we made a rough

adjustment for this factor (by subtracting the mean at a student's school from that student's score), the differences in means between Whites and Blacks (and Whites and Hispanics) were generally reduced by about 0.25 standard deviation units (but were still significant). This adjustment did not affect differences among groups in teacher ratings (because teachers were asked to base their ratings on how well a student performed relative to the other students in the same classroom).

TABLE 6
Mean Score for Whites Minus Minority Group Mean in Z-Score Units for Each Measure by Grade Level

	White–Hispanic mean			White–Black mean			White–Asian mean		
Measure	Grade five	Grade six	Grade nine	Grade five	Grade six	Grade nine	Grade five	Grade six	Grade nine
Teacher rating	0.49*	0.55*	0.28*	0.86*	0.40*	0.60*	0.12	−0.11	−0.52*
ITBS science	1.05*	1.03*	—	0.92*	1.09*	—	0.21	0.55*	—
CLAS other	0.89*	0.77*	0.73*	0.87*	0.93*	0.85*	0.14	0.31*	0.19
CLAS hands-on	0.96*	0.91*	0.61*	1.17*	0.96*	0.51*	0.24*	0.26*	−0.12
Incline	0.84*	—	—	1.13*	—	—	0.06	—	—
Friction	0.89*	—	—	0.80*	—	—	−0.02	—	—
Classification	—	0.72*	—	—	0.77*	—	—	0.27*	—
Inference	—	0.79*	—	—	1.15*	—	—	0.44*	—
Radiation	—	—	0.67*	—	—	0.86*	—	—	−0.12
Rate of cooling	—	—	0.65*	—	—	0.74*	—	—	−0.05

*Difference significant at $p < .05$. The CLAS hands-on score was based on three tasks. The score on each of the other six hands-on investigations was based on two tasks apiece.

An inspection of the inter-correlation matrix for each of the nine combinations of racial/ethnic group and grade level indicates that the strength of the relationship between two measures for Whites is very comparable to the relationship between them among Hispanics, Blacks, and Asians. Table 7 displays the typical pattern. These findings suggest that the constructs measured by a task (or test) are comparable across groups.

TABLE 7
Correlations Among Measures at Grade Six for Whites (Above the Diagonal) and Hispanics (Below the Diagonal)

	Teacher ratings	ITBS science	CLAS other	CLAS hands-on	Inference	Classification
Teacher ratings		0.46	0.45	0.37	0.45	0.44
ITBS science	0.41		0.59	0.48	0.50	0.48
CLAS other	0.44	0.57		0.44	0.48	0.49
CLAS hands-on	0.37	0.52	0.50		0.52	0.45
Inference	0.47	0.54	0.56	0.63		0.49
Classification	0.38	0.40	0.48	0.44	0.48	

Correlation of Performance Assessments with ITBS and Teacher Ratings

In all student groups studied, performance assessment scores correlated slightly (but not always significantly) higher with ITBS science scores than they did with teacher ratings. This trend appeared to stem from the following two factors—the rating a student received was made relative to the other students in the same classroom, whereas the students' score on a measure indicated how well that student performed relative to all of

the other students in the study and there was considerable variation among classrooms and schools in the average ability of their students. When we adjusted for this variation (by subtracting the classroom mean from each student's score in that classroom), the correlation between a hands-on and a multiple-choice test score was essentially the same size as the correlation of that hands-on measure with the teacher's ratings.

Differential Item Functioning

As indicated previously, we used shells to construct pairs of investigations (except those developed by CLAS) that included parallel tasks (as in the case of friction and incline or pendulums and levers). Each task had five or more separately scored questions or components. This design allowed us to examine whether size and direction of the difference in mean scores between tow groups of students on one question was consistent with the size and direction of the difference between these groups on the corresponding question on the other task that was developed from the same shell. For example, we could investigate whether the gender difference on the "interpretation" question in the friction investigation was the same as the gender difference on the interpretation question in the incline investigation.

The first step in these analyses involved converting the raw scores on each question within a task to z-scores. Next, we computed the difference in mean z-scores between groups on each question. We then used the question as the unit of analysis to compute the correlation between tasks within a shell (e.g., one pair of observations consisted of the student's score on the friction-interpretation question and that student's score on the incline-interpretation question).

These analyses suggested that certain questions produced larger differences between groups than other questions. However, further analyses revealed that with respect to race and ethnicity, virtually all of these relationships stemmed from some questions having much larger variances than others (usually because they had more scoreable components). Specifically, the larger the White's standard error on a question, the larger the White/Black or White/Hispanic difference in mean scores on that question. In short, what appeared to be differential item functioning was simply an artifact of some questions have more variance than others.

That was not the case with respect to gender. Regardless of the size of their standard errors, certain question types produced consistently larger differences in means between boys and girls than did other types. Using the question as the unit of analysis, the correlation between male-female differences in mean scores were 0.87 for friction and incline, 0.81 for pendulums and levers, 0.80 for animals and materials, and 0.49 for radiation with rate of cooling. Figure 3 illustrates this pattern. The data point in the upper right-hand corner of the figure shows that the question type that favored the girls the most over boys on the friction task corresponded to the same type of question that favored girls the most over boys on the incline task. On both investigations, the girls' mean on this question type was roughly 0.2 standard deviation units than the boys' mean.

To further investigate the source of the gender differences, we formed the following four clusters of questions: those on which boys did better than girls, those on which boys and girls did equally well, those on which girls did slightly better than boys, and those on which girls did much better than boys. An examination of these clusters

suggested that girls tended to do better on questions that required making the correct interpretation of the observed results of the experiment (e.g., whether the angle of the incline affects the force needed to pull the truck), whereas boys did better on questions that involved making predictions (e.g., whether more force would be needed if two large marbles were put in the truck).

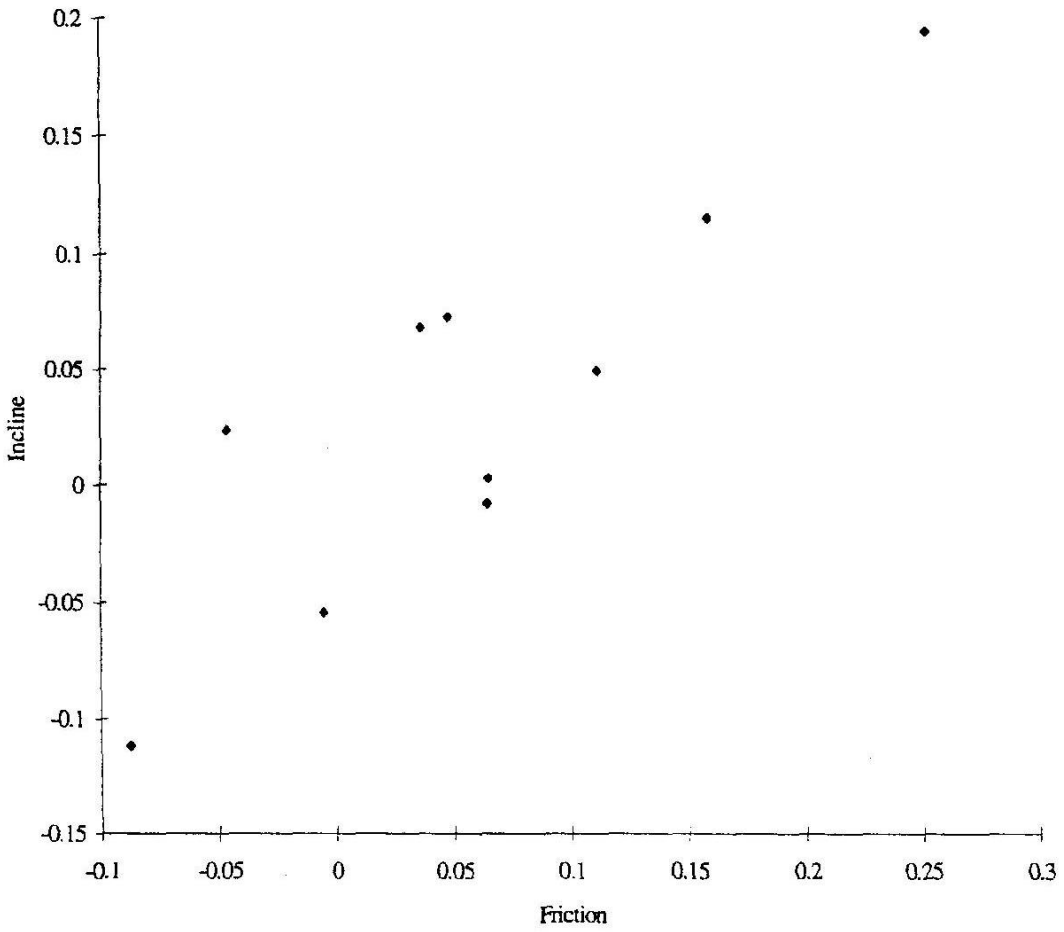

FIGURE 3. Girls' mean minus the boys' mean (in standard deviation units) on 10 corresponding questions on the incline and friction investigations.

Table 8 illustrates the differences among these question types by showing two types that favored girls and two that favored boys. Underlying the two types that favored girls was the ability to follow directions for conducting an experiment and to fill out the data-recording form accurately. Close attention to detail and instructions were important. In contrast, one of the question types that favored boys involved drawing inferences from the results to make a prediction about a condition that they could not test. The other pendulum-lever question that favored boys involved providing a rational for the use of a variable control.

TABLE 8
Corresponding Questions on Inference Tasks That Favored Girls or Boys

Question favors	Pendulums	Levers
Girls	(2) Which two pendulums took the most time to swing 20 times?	(2) Which two levers needed the most washers to lift the weight?
Girls	(3) What has the biggest effect on how fast a pendulum swings, its weight or its length?	(3) What has the biggest effect on a lever's ability to lift objects, its length or the location of its notch?
Boys	(4) How much time would it take Pendulum E to swing 20 times?	(4) How many washers would it take to lift the weight with Bar E?
Boys	(5) Was it important to have the line directly below the hook?	(5) Was it important for all the levers to have the same distance between the wooden "stop" and the end?

Note: Pendulum E and Bar E were the ones the student could observe but not test. The score on a question was a function of both the appropriateness of the answer and the rationale the student gave for that answer.

Conclusions

Girls tended to score slightly higher than boys on the performance assessments. However, although certain types of performance task questions favored girls, other types favored boys. It is not entirely clear why this happened, but we suspect it is related to the emphasis a question places on certain cognitive abilities or skill experience. Whatever the reason, our findings suggest that differences in mean scores between boys and girls on performance measures will be sensitive to the specific types of questions asked. Hence, subject matter is not the only factor that drives differences in scores between gender groups. Type of question within testing method also plays a role.

Our results regarding racial/ethnic differences are consistent with previous studies. Specifically, differences in mean scores among racial/ethnic groups were not related to test or question type. No matter which type was used, Whites had much higher means than Blacks or Hispanics. Thus, changing test or question type is unlikely to have much effect on the differences in mean scores among racial/ethnic groups. We found no empirical support for the hypothesis that because performance assessments involve "changing the game" from the familiar to the unfamiliar, they will be detrimental to minority groups (Baker & O'Neil, 1994).

Performance assessments typically require students to write short- to medium-length explanations of their answers. This may give girls an advantage (Baker, Freeman, & Clayton, 1991). Similarly, tasks that are especially sensitive to prior relevant knowledge may give White males and advantage over females and minority groups (Baker & O'Neil, 1994). In short, scores on performance assessments in science may be a function of differences in non-science abilities and experiences among groups. Of course, this also is true for traditional multiple-choice tests (Harmon, 1991). We anticipate these issues will be important considerations for those planning to use performance tasks in large-scale and high-stakes testing programs and for those charged with reporting and interpreting the results on these measures.

Notes

This material is based on work supported by the National Science Foundation under Grant No. MDR-9154406.

References

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitively sensitive assessment of subject matters: Understanding the marriage of psychological theory and educational policy in achievement testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). New York: Prentice Hall.

Baker, E. L., & O'Neil, H. F., Jr. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education, 1*(1), 11-25.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Education Measurement, 29*, 1-17.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 23-25.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165-174.

Bormuth, J. R. (171). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Carey, N. B., & Shavelson, R.J. (1989). Outcomes, achievement, participation, and attitudes. In R. J. Shavelson, L. M. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education* (pp. 147-191). Santa Monica, CA: RAND Corporation.

Erickson, G. L., & Erickson, L. J. (1984). Females and science achievement: Evidence, explanations, and implications. *Science Education, 68*, 63-69.

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39*(3), 193-202.

Glaser, R. (1998). Cognitive and environmental perspectives on assessing achievement. In E. E. Freeman (Ed.), *Assessment in the service of learning* (pp. 37-43). Princeton, NJ: Educational Testing Service.

Haladyna, T.M., & Shindoll, R. R. (1989). ISs: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions, 12*, 97-104.

Hanna, G. (1986). Sex differences in mathematics achievement of eighth graders in Ontario. *Journal of Research in Mathematics Education, 17*, 231-237.

Harmon, M. (1991). Fairness in testing. Are science education assessments biased? In G. Kulm & S. M. Malcolm (Eds.), *Science assessment in the service of reform* (pp. 31-54). Washington, CD: AAAS Publications.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1994). *Iowa Tests of Basic Skills: Interpretive guide for school administrators.* Chicago: Riverside.

Jenkins, L. B., & MacDonald, W. B. (1989). Science teaching in the spirit of science. *Issues in Science and Technology, 63*, 60-65.

Johnson, M. (1990). The science skills center, Brooklyn, NY: Assessing an accelerated science program for African-American and Hispanic elementary and junior high school students through advanced science examinations. In A. B. Champagne, B. E. Lovitts, & B. J. Calinger (Eds.), *Assessment in the service of instruction: This year in school science 1990* (pp. 103-126). Washington, DC: American Association for the Advancement of Science.

Johnson, S. (1987). Gender differences in science: Parallels in interest, experience, and performance. *International Journal of Science Education, 9*, 467-481.

Jones, L. R., Mullis, I. V., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). *The 1990 science report card: NAEP's assessment of fourth, eighth, and twelfth graders.* Princeton, NJ: Educational Testing Services.

Jovanovic, J., & Shavelson, R. J. (1995, April). *Examination of gender differences on performance-based assessments in science.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Journal of Education and Urban Society, 26*, 352-366.

Klein, S. P., Shavelson, R. J., Stecher, B. M., McCaffrey, D., Haertel, E., Comfort, K., Solano-Flores, G., & Jovanovic, J. (1996). *Sources of task sampling variability in science performance assessment tasks.* Manuscript in progress.

Langer, J. A., Applebee, A. N., Mullis, I. V. S., & Foertsch, M. A. (1990, June). *Learning to read in our nation's schools* (NAEP 19-R-02). Washington, DC: U.S. Department of Education.

Linn, M. C., Benedictis, T. D., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "I don't know" really mean? *Journal of Research in Science Teaching, 24*, 267-278.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991, November). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Mullis, I. V. S. (1994, October). *America's mathematics problem: Raising student achievement.* Washington, DC: U. S. Department of Education.

Murphy, R. J. L. (1982). Sex differences in objective test performances. *British Journal of Educational Psychology, 52*, 213-219.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Boston: Author.

Neil, D. M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 70*, 688-696.

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20*(5), 12-20.

Peng, S. S., Wright, D., & Hill, S. T. (1995). *Understanding racial-ethnic differences in secondary school science and mathematics achievement* (NCES 95-710). Washington, DC: U. S. Department of Education.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995, January). *Listening to children read aloud.* Washington, DC: US Department of Education.

Raizen, S., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. N. V., & Oakes, J. (1989). *Assessment in elementary school science education.* Washington, CD: National Center for Improving Science Education.

Rowley, G. L. (1974). Which examinees are most favored by the use of multiple choice tests? *Journal of Educational Measurement, 11*, 15-23.

Shavelson, R. J. (1997). *On the development of science performance assessment technology.* Manuscript in preparation.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*, 692-697.

Slatker, M. J. (1968). The effect of guessing strategy on objective test scores. *Journal of Educational Measurement, 5*, 217-221.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1997). *On the development and evaluation of a shell for generating science performance assessments.* Manuscript submitted for publication.

Stecher, B. M., & Klein, S. P. (Eds.). (1995). *Performance assessments in science: Hands-on tasks and scoring guides.* Santa Monica, CA: RAND.

Tamir, P. (1993). A focus on student assessment. *Journal of Research in Science Teaching, 30*, 535-536.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55-66.

Authors

Stephen P. Klein is a senior research scientist at RAND, 1700 Main Street, Santa Monica, CA 90407-2138. His specialty is educational measurement.

Jasna Jovanovic is an assistant professor of human development at the University of Illinois, Urbana-Champaign, Department of Human and Community Development, 1105 West Nevada Street, Urbana, IL 61801. She specializes in gender and achievement.

Brian M. Stecher is a research scientist at RAND at the above address. His specialty is educational measurement.

Dan McCaffrey is a statistician at RAND at the above address. He specializes in educational measurement.

Richard J. Shavelson is dean of the School of Education at Stanford University, Stanford, CA 94305-3096. His specialty is educational measurement.

Edward Haertel is a professor at the School of Education at Stanford University at the above address. He specializes in educational measurement.

Guillermo Solano-Flores is a senior research associate at WesEd, 730 Harrison Street, San Francisco, CA 94107-1242. His specialty is educational measurement.

Kathy Comfort is a project director at WestEd at the above address. She specializes in science assessment and curriculum.