Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society, 26*(4), 352-366.

Performance-Based Assessments
Will Gender Differences in Science Achievement be Eliminated?

Jasna Jovanovic**,** Guillermo Solano-Flores, & Richard J. Shavelson

By the time children in the United States reach seventh grade, half declare no interest in science (Office of Science and Technology Policy, 1991). Among girls, this disinterest appears to be particularly pronounced (S. Johnson, 1987; Jones, Mullis, Raizen, Weiss, & Weston, 1992). At the same time, girls' and boys' performances on standardized tests of science achievement begin to diverge with girls falling behind boys. This fact is well supported by numerous large-scale studies such as the International Association for the Evaluation of Educational Achievement or IEA (1988), the National Assessment of Educational Progress (NAEP) 1970-1986 (Mullis & Jenkins, 1988), and the British Columbia Science Assessments (Bateson & Parsons-Chatman, 1989). In the science classroom, however, girls perform as well, or better than, boys (Maccoby & Jacklin, 1974). Therefore, standardized tests are thought to under-predict girls' science achievement (Linn, 1991). Although this gender disparity has been attributed to several factors, there is considerable concern that the difference may be an artifact of the method of measurement (Bateson & Parsons-Chatman, 1989; Bolger & Kellaghan, 1990). That is, there is something about the test itself that puts girls at a disadvantage. Girls' lower test scores, in turn, are thought to undermine their self-perceptions of competence, leading to their disinterest in science and eventual drop from the science "pipeline" (Oaks, 1990; Rosser et al., 1989).

Included in the current reform rhetoric is the need to change the method by which we evaluate students' achievements. To do so will "open gates of opportunity rather than close them off" (National Commission on Testing and Public Policy, 1990, p. x). The belief seems to be that by replacing traditional assessment methods with new alternative methods such as performance-based assessments, the gender bias in testing may be eliminated (Jenkins & MacDonald, 1989; National Center for Improving Science Education, 1989). This article begins by reviewing what we know about gender differences on traditional tests of science achievement and what is hoped to be gained by changing to performance-based assessments. Then, as an initial look at the effect of new forms of testing on males' and females' science achievement, their scores on performance-based assessments are compared. Finally, these findings are discussed in the context of science reform.

Gender Differences on Traditional Science Achievement Tests

Conventional science achievement tests, like most achievement tests, are typically multiple-choice tests. When this method of measurement is used, females tend to perform

more poorly than males (Harding, 1981; Murphy, 1982). The reason for this female disadvantage is unclear. However, there is no evidence that multiple-choice tests penalize examinees who are reluctant to guess while favoring examinees who are test-wise and willing to take risks (Rowley, 1974; Slatker, 1968). In general, males are more willing than are females to take risks and to guess on multiple-choice tests (Hanna, 1986). Moreover, females are more willing than are males to choose an "I don't know" response option; a response that is not rewarded on standardized tests (Linn, Benedictis, Delucchi, Harris, & Stage, 1987). This becomes especially apparent on tests in specific content domains of science. For example, on physical science items there are large differences between males and females in "I don't know" responses (Linn, et al., 1987).

One interpretation of this response style among females is that they simply do not know the content area of physical science as well as males. Analysis of science achievement tests, item by item, indicates that the larges performance discrepancies occur on test items dependent on physics knowledge (S. Johnson, & Murphy, 1984). For example, in an examination of the British Columbia Science Assessment, Erickson and Erickson (1984) noted that males outperformed females on items related to principles of gravitation, electricity, and motion. As Erickson and Erickson (1984) postulate, males do better than females on items that deal with objects and events drawn from their "sphere of experience." Put another way, one possible explanation of females' tendency to select "I don't know" alternatives is due to their disadvantage in domains in science where they have fewer science-related experiences (S. Johnson, 1987; Kahle & Lakes, 1983). This possible explanation is supported by the 1990 NAEP survey of student attitudes at Grades 4, 8, and 12. Significantly more males than females reported experiences with activities related to physical science such as experimenting with batteries and magnets (Jones et al., 1992). In several studies, this "experiential" difference has been found to correspond to the discrepancies in male and female test scores in science (Erickson & Erickson, 1984; S. Johnson, 1987).

The emphasis of traditional achievement tests on the recall of basic content is thought to put females at a further disadvantage (Champagne & Newell, 1992). For example, on the NAEP 1976-1977 and 1981-1982 assessments, there were consistent gender differences among 13- and 17-year-olds on science items that stressed specific knowledge in particular content areas (e. g., "What is evaporation?"). On the other hand, gender differences were absent on items that involved analytic processes and multi-step reasoning such as designing experiments and drawing conclusions (Linn et al., 1987). Similarly, in the British Columbia Science Assessment at Grades 4, 8, and 12 on items requiring specialized content knowledge males outscored the females, whereas no differences emerged in understanding scientific processes (Erickson & Erickson, 1984). Apparently, the "female disadvantage" on science achievement tests vanishes when emphasis is place on problem solving, reasoning, and critical thinking.

Changing to Performance-Based Assessments

The impetus for change in science testing comes from the notion that measurement of content knowledge at the exclusion of process and application skills gives an incomplete picture of students' science achievement (Frederiksen, 1984; Glaser, 1988). Moreover, there is a belief that continued reliance on multiple-choice tests may

hamper the development of innovative learning technologies (Wiggins, 1989) and efforts to successfully implement the reform agenda (Shavelson, Carey, & Webb, 1990). To support science education reform and to reflect student achievement accurately, "all tests should involve students in the actual challenges, standards, and habits needed for success in the academic disciplines or in the workplace: conducting original research, analyzing the research of others, arguing critically, and synthesizing divergent viewpoints" (Wiggins, 1989, p. 706).

Performance-based science assessments provide students with hands-on opportunities to demonstrate their knowledge, not simply by recalling scientific facts but by constructing solutions (Baxter, 1991). In evaluating student performance thee is an emphasis on the process by which students generate solutions, not just on the correctness of the solution itself (Baxter, Shavelson, Goldman, & Pine, 1992; Carey & Shavelson, 1989). The notion is that individuals approach problem solving differently due to varying styles, not differing abilities (Paris, Lawton, Turner, & Roth, 1991). Therefore, tests should be constructed to allow for this individual variation (Neil & Medina, 1989).

As a consequence of performance-based testing's greater emphasis on problem solving and critical thinking, the idea is that factors such as guessing or reliance on exposure to science-related activities outside the classroom should be substantially reduced, giving way to the opportunity to females to show what they know and can do (Jenkins & MacDonald, 1989; M. Johnson, 1990). This proposition, however, has not been empirically explored. Moreover, critics of reform caution that too often new programs or methods are implemented without evidence that their impact on students (Slavin, 1989). Research and development are therefore needed to examine whether alternative approaches to science testing will be "fair" in their assessments of both males and females (Kulm & Stuessy, 1991).

A Preliminary Investigation

Over the past 5 years, a team of researchers and science teachers at the University of California, Santa Barbara and the California Institute of Technology have been developing and evaluating hands-on performance of measures in science (e.g., Shavelson, Baxter, & Gao, 1993; Shavelson, Baxter, & Pine, 1991, 1992; Solano-Flores, Jovanovic, & Shavelson, 1994). They have investigated the reliability and validity of performance assessment development for research and for the purposes of applied statewide performance assessment programs. Using data generated from these studies, the present investigation examined gender differences in these data sets. Table 1 includes the list of studies from which the science achievement data sets were generated, the investigators, sample sizes, percentage of females in each sample and grade levels.

TABLE 1
Science Achievement Data

Study	*Investigators*	N	*% Female*	*Grade*
A	Shavelson, Baxter, and Gao (1993); Gao (1992)	600	53	6
B	Baxter (1991); Shavelson, Baxter, and Pine (1992)	197	46	5, 6
C	Solano-Flores, Jovanovic, and Shavelson (1994)	109	51	5

As will be described below, all three studies involved the generation of science achievement data in several content areas in science. In addition, Study B assessed each content area through the use of several measurement methods (i.e., hands-on, short-answer, and multiple-choice). In Study C, assessments were composed of several science tasks (i.e., planning and design, performance, analysis and interpretation, and application). For all three data sets gender effect sizes were computed by subtracting the standardized mean score for males from the standardized mean score for females (Bolger & Kellaghan, 1990). In this way, positive values reflect a male advantage. In addition, to determine the significant effects of gender and its interaction with science content area, measurement method, and/or science behavior, a repeated analysis of variance (ANOVA) of the standard deviate achievement scores was applied to each data set. [1]

Study A

In Study A, science achievement scores were examined from a random sample of students who participated in the 1990 Science Performance Field Test of the California Assessment Program or CAP (Comfort, 1991; the name has recently been changed to the California Learning Assessment System). The Cap field test was composed of five hands-on investigations from five content areas in elementary school science:

- Electricity: Determine the conductivity of different materials.
- Leaves: Develop a classification system for leaves and then classify a new mystery leave.
- Rocks: Test and identify the properties of three rocks and then identify and unknown rock.
- Measurement: Select and appropriate measurement tool to estimate the quantities of height, volume, temperature, and mass.
- Acids and Bases: Determine the pH of water samples.

Individual students rotated through five self-contained stations to perform each investigation at timed intervals (about 13 minutes per station). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table). Performance on each investigation was scored holistically (for scoring and psychometric details, see Gao, 1992; Shavelson, Baxter, & Gao, 1993).

Figure 1 presents the gender effect sizes for science achievement for each of the five CAP investigations (i.e., five science content areas). A Gender x Content Area repeated measures ANOVA indicated a significant between-subjects effect for Gender,

$F(1, 593) = 4.87$, $p<.05$; and a significant Gender x Content Area interaction, $F(4, 2372) = 9.50$, $p<.0001$. Simple pair-wise comparisons of the three differences between means[2] indicated that females performed better than males on three of the five investigations: Leaves, $t(595) = -4.39$, $p<.05$; Rocks, $t(595) = -2.59$, $p<.05$; and Acids and Bases $t(595) = -2.07$, $p<.05$. No significant difference was found between males and females on the Measurement investigation, $t(595) = -1.01$, $p<.31$. On the Electricity investigation, a significant male advantage was found, $t(595) = 2.29$, $p<.05$.

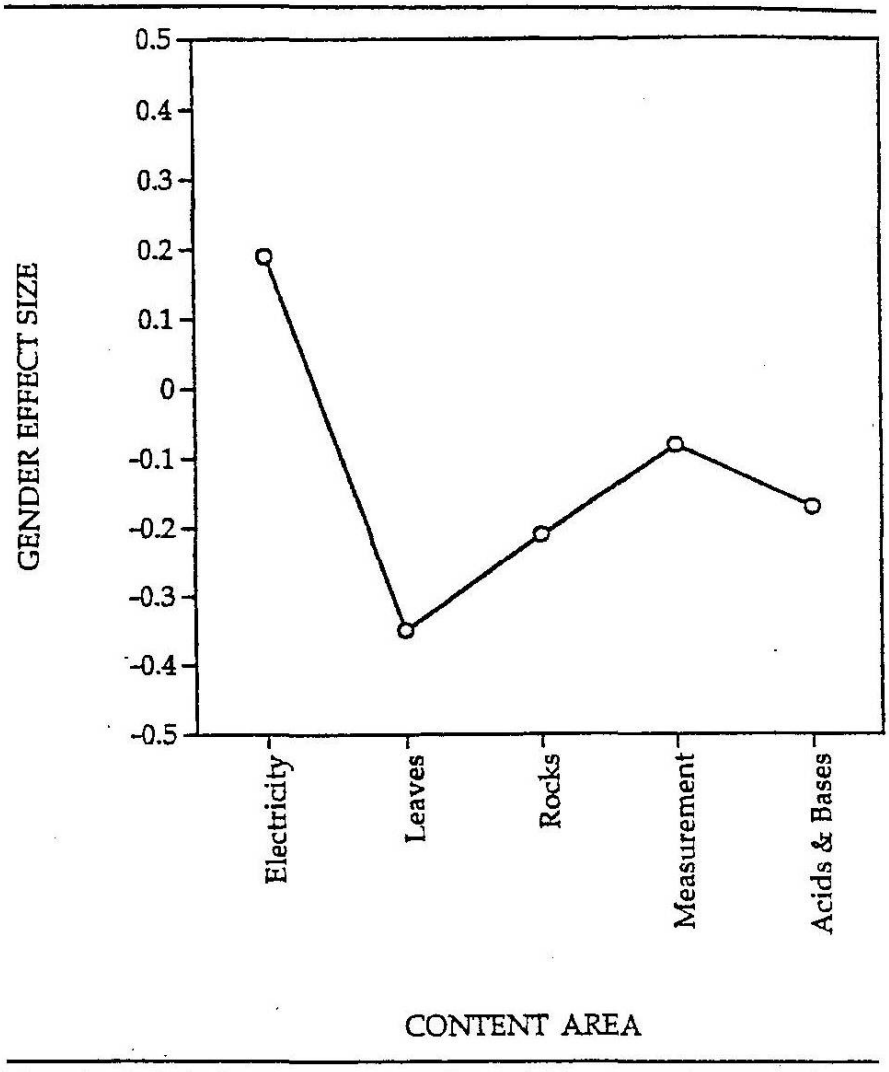

Figure 1: Study A: Gender Effect Sizes for Science Achievement by Content Area[a]
a. Positive value reflects male advantage.

Study B

In Study B, data were collected with three hands-on investigations from three content areas in science:

- Electric Mysteries: Determine the contents of six "mystery" boxes by connecting circuits to them.
- Bugs: Determine sow bugs' preferences for various environments (e.g., dark or light, dry or wet).
- Paper Towels: Determine which of three different paper towels soak up the most/least water.

Students' performances on each investigation were observed and scored by trained raters. Both short-answer questions and multiple-choice questions were developed to parallel the content of the three hands-on investigations. These questions were administered to each student at different times in the school semester prior to the hands-on investigation.[3]

Figure 2 shows the gender effect sizes for science achievement by content area and measurement method. A Gender x Content Area x Measurement Method repeated measures ANOVA indicated no overall effect for Gender $F(1, 195) = 3.10$, $p<.08$; no Gender x Content Area x Measurement Method interaction, $F(4, 780) = 1.04$, $p<.38$; and no Gender x Measurement Method interaction, $F(2, 390) = 0.681$, $p<.50$. However, a significant Gender x Content Area interaction was found, $F(2, 390) = 4.96$, $p< .01$. Simple pair-wise comparisons of the differences between means indicated that this interaction is reflected in the higher performance of males relative to females on the Electric Mysteries investigation, $t(197) = 3.06$, $p<.05$.

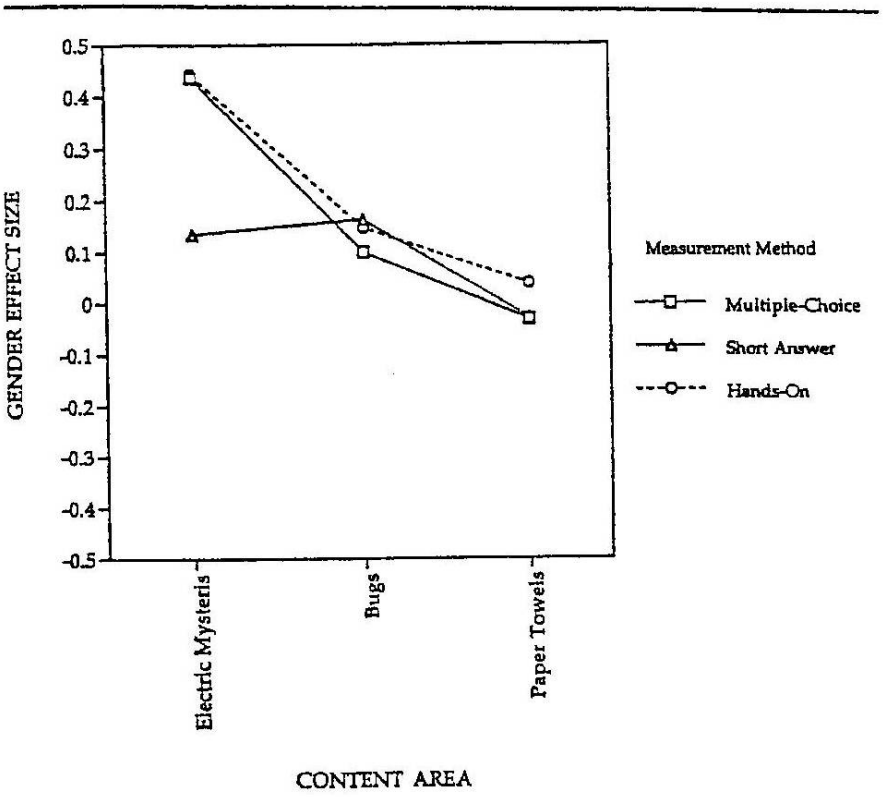

Figure 2: Study B: Gender Effect Sizes for Science Achievement by Content Area and Measurement Method[a]

a. Positive value reflects male advantage.

Study C

In Study C, data were collected with two performance-based investigations that sampled two content areas from the concept domain of force and motion:

- Incline Planes: Determine the relationship between the slope of a plane and the force needed to move an object to the top of the plane.
- Friction: Determine the relationship between surface texture and the force needed to move an object on the surface.

Each investigation was composed of four science tasks relevant to an inquiry-oriented approach to science (Tamir, 1974):

- Planning and design: conceiving a procedure to experimentally test a hypothesis
- Performance: conducting an experiment to test a hypothesis
- Analysis and interpretation: transforming and representing information and inferring
- Application: solving a new, concrete problem by using scientific principles used in the experiment.

Figure 3 presents the gender effect sizes for achievement by content area and science task. A Gender x Content Area x Task repeated measures ANOVA indicated no overall Gender effect, $F(1, 99) = .19$, $p<.66$; no Gender x Content Area x Task interaction, $F(3, 297) = .19$, $p<.90$; no Gender x Content Area interaction, $F(1, 99) = .14$, $p<.71$; and no Gender x Task interaction, $F(3, 297) = .01$, $p<.99$.

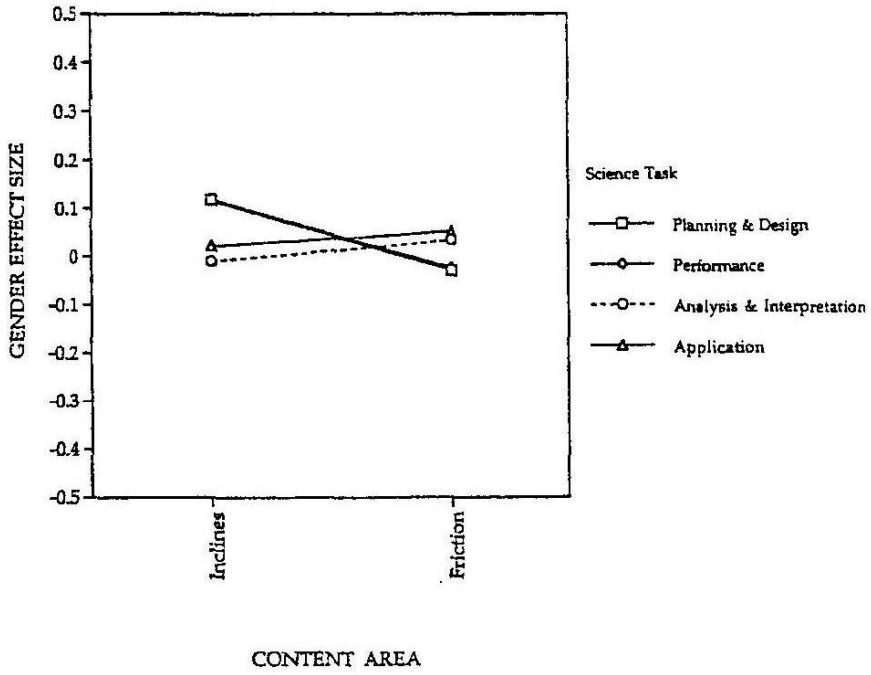

Figure 3: Study C: Gender Effect Sizes for Science Achievement by Content Area and Science Task[a]

a. Positive value reflects male advantage.

Interpretation and Discussion

Overall, the picture that emerges from this preliminary investigation is that there are few differences between males and females on performance-based assessments at the elementary school level. Summing across studies, science achievement on 5 of the 10 performance-based assessments showed no gender effect. Differences that did emerge appeared to be dependent on the particular science content area that was assessed. Females performed better than males on activities related to the classification of leaves and rocks (.35 and .21 standard deviation units higher relative to males, respectively) and somewhat better on the activity involving the measurement of the pH levels of water samples (.17 standard deviation units higher relative to males). Males, on the other hand, had a persistent advantage over females on activities related to electricity. Moreover, this male advantage resulted despite the method of measurement; across all three studies, the largest gender effect was found for the Electric Mysteries hands-on assessment (i.e., males performed .44 standard deviation units higher relative to females).

The present results suggest that, regardless of the method of measurement, students' prior experiences play a role in their testing performances. As previously mentioned, in general, boys' experiment with batteries and bulbs more often that girls (Jones et al., 1992), whereas girls collect flowers and/or plants more often than boys (S. Johnson, 1987). These differential experiences reported in previous research may account, in part, for the performance differences in the present results. It is unlikely, however, that prior experience alone accounts for the gender differences that were found. Clearly, students are situated in classrooms that are instrumental in their learning of science. Ideally, within the classroom, boys and girls should have the same instructional experiences. This, however, is not necessarily the case. For example, certain instructional practices (Eccles & Blumenfeld, 1985) or teacher-student interactions may favor one gender or another (Brophy & Good, 1974; Tobin, Kahle, & Faser, 1990). At the same time, male and female students' prior experiences may interact with instruction differently (Good & Stipek, 1983; S. Johnson & Murphy, 1984). For example, if a student has prior experience with a particular content area (e.g., connecting batteries to bulbs), he or she may be more willing to respond to the teachers' questions regarding the content area (e.g., electricity) or to take and active role in an activity that focuses on the content area (e.g., constructing an electric circuit). Furthermore, the teacher's preconceptions about gender differences in particular content domains (e.g., that males are better at physical science, or that girls are better at biology) may lead the teacher to call on, or to respond to, males and females differently (Morse & Handley, 1985). This prior-experience-by-instruction interaction may lead to varying learning experiences for males and females. These differences may, in turn, be reflected in students' test performances.

As outlined in a policy report issued by the National Center for Improving Science Education, the goals of alternative science assessments are to

(1) provide greater opportunities for children to interact with stimulus materials, (2) attend to understandings of constructs and principles as well as factual knowledge, (3) probe approaches to problem solving as well as outcomes, (4) be explicitly integrated with the curriculum and with instruction, (5) incorporate hands-on activities whenever feasible, and (6) be structured around group as well as individual activities. (Raizen et al., 1989, p. 97).

If teachers are held accountable to such goals, then the focus of instruction in the science classroom will change (Champagne & Newell, 1992), That is, tests that focus on a range of science skills will motivate teachers to spend time helping *all* students to develop these skills (Shavelson et al., 1992). In this way, the classroom environment may become an "equalizer," providing compensation for the disparities in students' experiences outside of school (Jenkins & MacDonald, 1989).

An important issue that could not be addressed in the present investigation is how performance-based testing will affect males' and females' attitudes and interests in science. Research on achievement motivation indicates that repeated experiences of failure diminish students' perceptions of their ability (Diener & Dweck, 1978; Harter, 1985). When students feel that their achievement failures are caused by low ability, these feelings can lead to shame, humiliation, and anxiety in evaluative situations (Weiner,

1984). In turn, low-achievement scores may decrease motivation to learn (Diener & Dweck, 1978). Therefore, the issue of whether performance-based assessments (and the corresponding performance-based curricula) will facilitate or impede students' learning in science must become an integral part of the evaluation of these new tests.

Clearly, tests in themselves are not going to eliminate gender difference in science achievement; the issue is far more complicated. But given the major influence tests can have (e.g., determining scholarships and college entrance), educational reformers must be sensitive to the differential impact that alternative assessments can have on male and/or female performance in learning science. There is a danger in simply assuming that performance-based assessments will be fair in their treatment of females (or, for that matter, other minority groups in science). The danger is that, if girls do not do better, this may only further the popular cultural stereotype that girls simply cannot do science as well as boys (Kelly, 1988). Steps have been taken to ensure that all students, regardless of gender, acquire the knowledge and skills necessary for successful achievement in science.

Authors' note: *This research was supported, in part, by grants to Richard J. Shavelson from the California Assessment Program, the National Science Foundation (No. TPE-90-55443), and the RAND Corporation (No. 92-13/I). The views expressed herein do not necessarily reflect those of the sponsoring agencies. The authors wish to thank Gail Baxter and Xiahong Gao for their major contributions to the generation of two of the data sets used in the present investigation. Thanks are also due to Michelle Perry and Clare Stocker for their feedback on an earlier version of this article. Correspondence and requests for reprints should be sent to Jasna Jovanovic, University of Illinois, U-C Division of Human Development and Family Studies, 1105 W. Nevada, Urbana, IL 61801.*

---

[1] In Study B, for example, a 2 x 3 x 3 (Gender x Content Area x Measurement Method) analysis of variance (ANOVA) with repeated measures on the Content Area and Measurement Method factors was carried out. Because the effects of gender and not the main effects of Content Area and Measurement Method were of interest in the present investigation, scores on the nine measures of science achievement were converted to z scores to account for scale variation (Bolger & Kellaghan, 1990).

[2] Given the exploratory nature of the present investigation, each comparison was carried out at an alpha level of .05. The intent was to avoid a Type II error—incorrectly concluding no gender difference. An alternative, dividing the alpha level by the number of comparisons, would protect against a Type I error—incorrectly concluding a gender difference. The former was not as of great concern here.

[3] In addition to the short-answer and multiple-choice "surrogates," a computer simulation for the Electric Mysteries and Bugs investigations was also administered (for a description, see Baxter & Shavelson, in press). However, because this method was not developed for the Paper Towel investigation, it was dropped from the present analysis.

References

Bateson, D. J., & Parsons-Chatman, S. (1989). Sex-related differences in science achievement: A possible testing artifact. *International Journal of Science Education, 11*, 371-385.

Baxter, G. P. (1991). *Exchangeability of science-performance assessments.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Baxter, G. P., & Shavelson, R. J. (in press). Science performance assessments: Benchmarks and surrogates. *International Journal of Education.*

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

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

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences.* New York: Holt, Rinehart & Winston.

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

Champagne, A. B., & Newell, S. T. (1992). Directions for research and development: Alternative methods of assessing scientific literacy. *Journal of Research in Science Teaching, 29*, 841-860.

Comfort, K. B. (1991). A national standing ovation for the new performance testing. In G. Kulm & S. M. Malcolm (Eds.), *Science assessment in the service of reform* (pp. 149-162). Washington, DC: American Association for the Advancement of Science.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance strategy, and achievement cognition's following failure. *Journal of Personality and Social Psychology, 36*, 451-462.

Eccles, J.S., & Blumenfeld, P. (1985). Classroom experiences and student gender: Are there differences and do they matter? In L. C. Wilkinson & C. B. Marett (Eds.), *Gender influences classroom interaction* (pp. 19-114). New York: Academic Press.

Erickson, G. L., & Erickson, L. J. (1984). Females and science achievement: Evidence, explanation, and implications. *Science Education, 68,* 63-89.

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

Gao, X. (1992). *Generalizability of a state-wide science performance assessment.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Glaser, R. (1988). Cognitive and environmental perspectives on assessing achievement. In E. E. Freeman (Ed.), *Assessment in the service of learning* (pp. 37-44). Princeton, NJ: Educational Testing Service.

Good, T. L., & Stipek, D. J. (1983). Individual differences in the classroom: A psychological perspective. In G. D. Fenstermacher & J. I . Goodland (Eds.), *Individual differences and the common curriculum: Eighty-second yearbook of the National Society for the Study of Education* (pp. 9-43). Chicago, IL: University of Chicago Press.

Hanna, G. (1986). Sex differences in science examinations. In A. Kelly (Ed.), *The missing half: Girls and science education* (pp. 192-204). Manchester, England: Manchester University Press.

Harter, S. (1985). Competence as a dimension of self-evaluation: Toward a comprehensive model of self-worth. In R. L. Leahy (Ed.), *The development of the self* (pp. 55-121). New York: Academic Press.

International Association for the Evaluation of Educational Achievement (1988). *Science achievement in seventeen countries: A preliminary report.* New York: Pergamon.

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. 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.

Johnson, S., & Murphy, P. (1984). The underachievement of girls in physics: Toward explanations. *European Journal of Science Education, 6*, 399-409.

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.

Kahle, J., & Lakes, J. (1983). The myth of equality in science classrooms. *Journal of Research in Science Teaching, 20*(2), 131-140.

Kelly, A. (1988). Sex stereotypes and school science: A three year follow-up. *Educational Studies, 14*(2), 151-163.

Kulm, G., & Stuessy, C. (1991). Assessment in science and mathematics education reform. In G. Kulm & S. M. Malcolm (Eds.), *Science assessment in the service of reform* (pp. 71-87). Washington, DC: American Association for the Advancement of Science.

Linn, M. C. (1991). Gender differences in educational achievement [Summary]. In *Proceedings of the 1991 ETS Invitational Conference: Sex equity in educational opportunity, achievement, and testing* (pp. 11-50). Princeton, NJ: Educational Testing Services.

Linn, M. C., Benedicitis, 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.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Morse, L. W., & Handley, H. M. (1985). Listening to adolescents: Gender differences in science classroom interaction. In L. C. Wilkinson & C. B. Marett (Eds.), *Gender influences in classroom interaction* (pp. 37-56). New York: Academic Press.

Mullis, I.V. S., & Jenkins, L. B. (1988). *The science report card: Elements of risk and recovery.* Princeton, NJ: Educational Testing Services.

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

National Center for Improving Science Education. (1989). *Getting started in science: A blueprint for elementary school science education.* Andover, MA: The NETWORK, Inc.

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

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

Oakes, J. (1990). *Lost talent: The under-participation of women, minorities, and disabled person in science.* Santa Monica, CA: RAND.

Office of Science and Technology Policy. (1991). *By the year 2000: Report of the FCCSET Committee on Education and Human Resources.* Washington, DC: Author.

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.

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

Rosser, P., Brown, S., Greenberger, M., Johnson, S., Medaus, G., Welsh, M., & Wolfe, L. (1989). Gender bias in testing: Current debates for future priorities. A public policy dialogue. *Proceedings of the Ford Foundation Women's Program Forum.* New York: Ford Foundation.

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., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*, 347-362.

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

Slavin, R. E. (1989). PET and pendulum: Faddism in education and how to stop it. *Phi Delta Kappan, 70*, 752-758.

Solano-Flores, G., Jovanovic, J., & Shavelson, R. J. (1994, April). *Development of an item shell for the generation of performance assessments in physics.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Tamir, P. (1974). An inquiry oriented laboratory examination. *Journal of Educational Measurement, 11*, 25-33.

Tobin, K., Kahle, J. B., & Fraser, B. J. (1990). *Windows into science classrooms: Problems associated with higher-level cognitive learning.* New York: Falmer.

Weiner, B. (1984). Principles for a theory of student motivation ant their application within and attributional framework. In R. Ames & C. Ames (Eds.), *Research on motivation in education* (Vol. 1, pp. 15-38). New York: Academic Press.

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