Exploring Differential Item Functioning on Science Achievement Tests

CSE Technical Report 483

Laura S. Hamilton, RAND
Richard E. Snow, CRESST/Stanford University

August 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences. Richard Shavelson, Project Director, CRESST/Stanford University

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of he National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

This research was also supported in part by a Spencer Foundation Dissertation Fellowship and National Science Foundation grant RED-9253068. Parts of the study were also supported by the American Educational Research Association, which receives funds for its AERA Grants Program from the National Science Foundation and the National Center for Education Statistics (U.S. Department of Education) under NSF Grant #RED-9452861. Opinions reflect those of the authors and do not necessarily reflect those of these granting agencies.

EXPLORING DIFFERENTIAL ITEM FUNCTIONING ON SCIENCE ACHIEVEMENT TESTS[1]

Laura S. Hamilton, RAND
Richard E. Snow, CRESST/Stanford University

Abstract

This study explores methods for detecting gender-based differential item functioning (DIF) on 12th grade multiple-choice and constructed-response science tests administered as part of the National Education Longitudinal Study of 1988 (NELS:88). Several combinations of conditioning variables were explored for DIF detection on both tests, and results were supplemented with evidence from interviews of students who completed the test items. On both tests, DIF in favor of males was exhibited primarily on items that involved visualization and that called upon knowledge and experiences acquired outside of school. The findings revealed that neither content nor format alone explained the patterns of male and female performance, and that an investigation of response processes may provide valuable additional information about the nature of gender differences in science achievement.

Assessment in the United States has served a variety of purposes including instructional feedback, selection, and assignment to educational programs. Educational reforms that have taken place during the past three decades, as well as concerns about achievement and equity, have given rise to an increase in the use of assessment for purposes of monitoring and accountability (Linn, 1989). Assertions about the need for assessment to monitor achievement can be found in such documents as the widely-publicized *A Nation at Risk* (National Commission on Excellence in Education, 1983) and in more recent reports such as that by the National Council on Education Standards and Testing (1992). It has therefore been necessary to develop assessments that can be administered to large samples of students across the nation.

Because of the need for standardization and inexpensive scoring, most large-scale testing programs have relied on the multiple-choice (MC) item format. In recent years, however, many assessment programs have adopted

[1]We are grateful to Vi-Nhuan Le and Judy Dauberman for their assistance in interviewing and scoring.

open-ended item formats to supplement or replace MC items. Open-ended items are often presumed to measure reasoning in a way that is difficult or impossible with the MC format (Frederiksen, 1984; Resnick & Resnick, 1992; Shavelson, Carey, & Webb, 1990). To support these claims for any given test, a careful validity investigation must be carried out (Messick, 1989).

One of the presumed benefits of CR (constructed response) items, particularly on science tests, is a reduction in gender differences. Studies have revealed small but potentially important differences in the average measured science achievement of males and females (see, for example, Jones, Mullis, Raizen, Weiss, & Weston, 1992), and some evidence suggests that in fact such differences are larger on MC than on CR assessments (Bolger & Kellaghan, 1990; Mazzeo, Schmitt, & Bleistein, 1993). However, results are inconsistent, with open-ended items sometimes showing larger differences (e.g., Dunbar, Koretz, & Hoover, 1991; Mullis, Dossey, Owen, & Phillips, 1991). Furthermore, a recent review and synthesis conducted by the Educational Testing Service revealed no clear format effect (Cole, 1997).

In contrast, there have been fairly consistent findings with regard to the effect of content on gender differences in science achievement. Males, on average, outperform females on physical science items, whereas little or no difference is typically observed on life science items (Becker, 1989; Burkam, Lee, & Smerdon, 1997; Fleming & Malone, 1983; Jovanovic, Solano-Flores, & Shavelson, 1994; Young & Fraser, 1994). On the 1991 International Assessment of Educational Progress (IAEP), the largest male advantage occurred for physical science and earth and space science items (Beller & Gafni, 1996). Some studies have traced such differences to course-taking patterns or other aspects of opportunity to learn, including participation in extracurricular activities related to science (Johnson, 1987; Linn, 1985; NAEP, 1988).

The type of reasoning elicited by different types of items may also affect the degree to which items exhibit gender differences. In particular, males tend to outperform females on measures requiring visual or spatial processing (Halpern, 1997; Lohman, 1993).Although the implications of this difference for achievement in science have not been explored extensively, there is some evidence that it affects performance on certain types of mathematics items (e.g., Fennema & Tartre, 1985; Halpern, 1992). Males tend to perform better on geometry items than do females who are matched on total test score (O'Neill &

McPeek, 1993), a result which may reflect the spatial demands of geometry. The male advantage in spatial skills may stem in part from differential exposure to activities that help to develop those skills (Halpern, 1992; Linn & Hyde, 1989). Careful study of the features of items exhibiting gender differences is needed to understand the complex relationships among format, content, and reasoning processes and their effects on the performance of males and females.

This investigation focuses on gender differences on multiple-choice and constructed-response science items administered as part of the National Education Longitudinal Study of 1988 (NELS:88). The research combines an exploratory differential item functioning (DIF) study with a small-scale interview study to provide evidence concerning sources of gender differences on the test items. The study also reveals ways in which the identification of items exhibiting DIF depends upon the conditioning variables used. Implications for users of large-scale achievement test data are discussed.

Methods for Detecting DIF

Indices of differential item functioning (DIF) reveal whether members of two groups, equated on the relevant ability, have different probabilities of answering a particular item correctly. Established procedures exist for dichotomous items (e.g., Angoff, 1993), and currently much work is being done to investigate DIF indices for polytomously scored items. For example, the generalized Mantel-Haenszel (GMH) statistic (Agresti, 1990; Somes, 1986) extends the commonly used Mantel-Haenszel (MH) procedure to items with more than two scoring categories (which are treated as unordered). Miller and Spray (1993) describe a logistic regression procedure that extends the logistic regression model described by Swaminathan and Rogers (1990). Miller and Spray also discuss a logistic discriminant function analysis (LDFA) procedure in which probabilities of group membership are predicted from item and total test scores.

A difficulty that frequently arises with open-ended tasks is the absence of a suitable matching criterion: Total score is often not feasible because of the small number if items administered on a typical performance assessment. A multiple-choice test in the same subject may be appropriate if the two formats tap similar abilities. However, if this assumption of construct equivalence between formats does not hold, DIF may be confused with item impact (differences in item performance due to differences in group means on a relevant ability) because

students are not matched on the ability being measured by the studied test (Welch & Miller, 1995).

For complex performance tasks, which may tap a number of abilities, a multivariate matching procedure may be most appropriate. Studies of test data as well as simulations have demonstrated that matching on more than one ability can substantially reduce the number of items identified as exhibiting DIF and can reduce the probability that item impact is misinterpreted as DIF (Ackerman, 1992; Mazor, Kanjee, & Clauser, 1995). Several studies have examined the effects of matching on multiple abilities in real and simulated data (e.g., Clauser, Nungester, Mazor, & Ripkey, 1996; Douglas, Roussos, & Stout, 1997). It is also possible to condition on both ability and an educational background variable (Clauser, Nungester, & Swaminathan, 1997; Zwick & Ercikan, 1989). Logistic regression, because it allows for multiple matching criteria, appears especially promising for the analysis of multidimensional data. The present study examines the effects of a variety of matching criteria on the number and types of items identified as exhibiting DIF on multiple-choice and constructed-response science achievement tests. The study is not designed to compare DIF detection procedures; several of the studies cited above include simulations that were carried out for this purpose. Instead, it is an exploration of the features of items that exhibit DIF in a set of actual science test data.

Design and Methodology

The NELS:88 HSES Sample and Science Tests

NELS:88, sponsored by the National Center for Education Statistics (NCES), is the most recent in a series of large-scale surveys designed to monitor the educational progress of the nation's students. NELS:88 followed a national probability sample of 8th graders into the 10th and 12th grades using a series of cognitive tests as well as questionnaires completed by students, parents, teachers, and school administrators. NCES conducted a supplementary study, called the High School Effects Study (HSES), in which 10th graders from 247 high schools were sampled in 1990 and followed into the 12th grade.

Students took four multiple-choice (MC) tests at each grade level, in math, science, reading, and history. The science test included 25 items at each grade, with a 20-minute time limit. Six of the 25 10th-grade items were dropped and

replaced with new items at Grade 12. All items were scored as correct or incorrect. At the 12th grade, a subsample of the HSES students completed constructed-response (CR) items in either math or science. The present study focuses on the science test. Four items were administered in science, each with a time limit of 10 minutes. The items required students to supply brief written answers including, in some cases, diagrams. Scorable records were obtained for 2204 students from 108 schools. Although the items are not presumed to cover the domain of 12th-grade science, the designers attempted to include items varying in content and format. Furthermore, the items were designed to be attempted even by students with limited science background. They included content that was presumably familiar to all students, so that most could answer portions of the item, but complete answers to all parts required fairly sophisticated knowledge. The four science items included: (1) Nuclear and Fossil Fuels (CR1; hereafter, "Fuels"): Write a brief essay outlining advantages and disadvantages of each; (2) Eclipses (CR2): Produce diagrams of solar and lunar eclipses and explain why one can be seen from a greater geographical area on earth; (3) Rabbit and Wolf Populations (CR3; hereafter, "Populations"): Given graph representing population of rabbits, produce graph representing population of wolves, subject to certain constraints, and explain features of graph; (4) Heating Curve (CR4): Explain segments of graph representing temperature of a mixture as a function of time (mixture contains water and ice, and is being heated over an open flame).

The CR items were scored by teams of readers, mostly high school science teachers. Each problem was broken down into components or features, scored using categories of possible responses (based on test developers' predictions and results of pilot work). This analytic scoring system preserved information on specific parts of students' responses. After scoring was completed, the readers and test developers created a system for combining the analytic scores into a set of ordered categories for each item. This process resulted in a six-point scale score for each item, with 0 representing an apparent absence of understanding and 5 representing complete and correct responses to all parts of the item. Interrater reliability was evaluated by NCES and found to be adequate; this issue is not discussed here. Additional information about the items and their scoring can be found in the NCES report by Pollack and Rock (1995).

Sample weights were provided for the entire HSES sample, but these are not appropriate for the sample used in this study. Because many factors are likely to

have contributed to school administrators' decisions to allow the constructed-response tests to be administered, the sample of schools used in this study cannot be considered a random sample of HSES schools. Furthermore, the processes governing participation of students in the CR study are unknown, precluding accurate adjustment for nonresponse. Therefore, weights are not used in the analyses reported here, and results should be interpreted in light of this fact.

Statistical Analysis

Previous studies of the NELS:88 MC math and science tests suggested that they should be treated as multidimensional (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). Therefore full information item factor analysis of the science multiple-choice items was conducted for the HSES sample, to study the similarity in structure with the NELS:88 core sample and to provide cognitive variables (factor scores) for use in other analyses. This method has been used extensively with dichotomous items (see Bock, Gibbons, & Muraki, 1988, for technical details). Separate analyses were conduced on the 10th- and 12th-grade science tests. Factor scores were computed for the resulting dimensions; these expected a posteriori (EAP) scores are Bayes estimates of the mean of the posterior ability distribution, given the observed response pattern. The decision concerning how many factors to retain was based on chi-square change criteria as well as on substantive interpretability. Promax rotation of factors was used.

Differential Item Functioning (DIF) detection methods were used to investigate gender differences on both the MC and CR tests. DIF procedures allow the researcher to discover whether equally able members of two groups have different probabilities of answering a given item correctly. The notion of "equally able" is one of the most problematic aspects of DIF detection. For MC tests, group members are generally equated on total test score and each item is studied separately. Of course this method cannot detect a bias that affects all items on the test. One of the primary difficulties in detecting DIF on CR items is the absence of a suitable matching criterion (also called a conditioning variable). This study explores ways in which DIF detection is influenced by changing the conditioning variable.

The first part of the DIF study focuses on the MC science test. Although these items had been tested for DIF prior to their inclusion on the test, it seemed worthwhile to explore the effects of treating the test as multidimensional rather than using total score alone. The Mantel-Haenszel (MH) chi-square method (Mantel & Haenszel, 1959), which is most commonly used for dichotomous items, was applied and compared with a logistic regression procedure. These two methods have been shown to function similarly in simulation studies (Swaminathan & Rogers, 1990). The latter was conducted three times for each item: once conditioning on total IRT score, once using the dimension on which the item loaded most highly, and once using all three dimensions. These analyses reveal ways in which definitions and interpretations of DIF change depending upon the conditioning variables used.

The logistic discriminant function analysis (LDFA) procedure is the primary method used to investigate DIF on the CR items (Miller & Spray, 1993). This method is more flexible than chi-square methods and has greater power to detect non-uniform DIF. Furthermore, for polytomously-scored items, most chi-square methods treat the response categories as unordered, resulting in a loss of information. In the LDFA procedure, probabilities of group membership (in this case, male vs. female) are predicted from total test score, item score, and their interaction, with likelihood ratio tests conducted for main effects and interaction models. Results of analyses using several sets of conditioning variables were compared, and both uniform and non-uniform DIF were investigated (the latter term refers to cases in which the magnitude of DIF varies by ability level). Conditioning variables included total CR score and total item response theory (IRT) score from the MC science test, as well as the science dimensions that emerged from the full information item factor analysis. As suggested by Miller and Spray (1993), for items that exhibited DIF, confidence bands were constructed around the estimated logistic discriminant function to assess the practical importance of DIF.

Interview Study Procedures[2]

To supplement the statistical analysis, 25 local high school students were interviewed and asked to think aloud as they completed the four CR items and a subset of 16 MC items. Participants also responded to a set of post-test interview

[2]Additional details about the methodology used in the interview study are available in Hamilton (1997).

questions that elicited additional information concerning solution strategies and sources of knowledge. Interviews were audiotaped and transcribed, and interviewers used a structured observation sheet to record events that would not be captured on audiotape, such as the use of gestures. The four CR items were scored by two raters using the rubrics provided by the item developers (Pollack & Rock, 1995). Agreement was adequate, with Kappa values of .61 for Nuclear and Fossil Fuels, .78 for Eclipses and Heating Curve, and .84 for Rabbit and Wolf Populations. A third rater scored the papers on which the original two raters disagreed, and the final score assigned to each paper was the one on which two of the three raters agreed.

Several coding categories were created for the 16 MC items. Coding categories were selected to capture a range of strategies for responding to MC items and to identify the most common sources of knowledge. Selection was based in part on observations gathered during a previous study (Hamilton, Nussbaum, & Snow, 1997) in which the SM items, especially, tended to evoke particular types of responses such as gestures and visualization. Codes were created separately for each CR item and tailored to particular item characteristics, based in part on results from the earlier interview study. Four transcripts were randomly selected for coding by a second rater; Kappa values were .80 or higher for each (these statistics were calculated for MC and CR together). Discrepancies were resolved through discussion.

Results of Statistical Analyses

In this section, results of the MC factor analyses are described. Means and frequencies for males and females on the MC dimensions and CR items are presented, along with correlations among measures of science achievement. DIF results for the MC and CR tests are discussed, with an emphasis on characteristics of items that contribute to measured gender differences.

Factor Analysis of MC Test

Results of the full information item factor analysis of the HSES MC data are presented in Table 1. At each grade, three dimensions emerged: Spatial-Mechanical Reasoning (SM), including items that required interpretation of visual or spatial relations; Quantitative Science (QS), involving chemistry and physics content and use of mathematical formulas; and Basic Knowledge and

Table 1

Factor Loadings from Full-Information Factor Analysis of NELS:88 10th- and 12th-Grade Science Multiple-Choice Test Items After Promax Rotation, HSES Sample (N = 5224-7191)

Master item number	Description	10th-grade						12th-grade					
		SM		QS		BKR	Comm. est.	SM		QS		BKR	Comm. est.
S27	Lever	0.64	*	0.17		-0.01	0.66	0.72	*	0.06		0.04	0.54
S29	Camera lens	0.71	*	-0.03		0.11	0.66	0.67	*	0.15		0.01	0.71
S28	Contour map	0.55	*	0.32		-0.06	0.58	0.56	*	0.14		0.20	0.67
S12	Earth orbit	0.30		-0.01		0.47	* 0.57	0.50	*	-0.13		0.41	0.59
S36	Pendulum							0.43	*	0.11		0.29	0.87
S14	Mix water	0.37	*	0.25		0.30	0.78	0.39	*	0.18		0.29	0.64
S38	Train							0.33	*	0.29		0.26	0.52
S37	Hydro. react.							0.13		0.77	*	-0.13	0.57
S35	Uranium decay							-0.10		0.77	*	0.14	0.70
S30	Half life	0.29		0.50	*	0.10	0.66	0.24		0.67	*	0.03	0.78
S26	Calc. mass	0.18		0.80	*	-0.11	0.76	0.23		0.56	*	0.17	0.83
S16	Enzyme graph	0.06		0.37	*	0.30	0.48	0.10		0.54	*	0.20	0.60
S05	Moon's light	0.20		-0.28		0.82	* 0.62	0.25		-0.26		0.72	* 0.59
S06	Simple reflex	-0.07		-0.05		0.88	* 0.67	0.08		-0.09		0.72	* 0.61
S17	Algae	0.28		0.15		0.41	* 0.65	0.16		0.06		0.66	* 0.74
S04	Expt. design	-0.09		0.13		0.48	* 0.32	-0.07		0.12		0.59	* 0.40
S10	Classify subs.	0.16		0.22		0.45	* 0.64	0.16		0.13		0.56	* 0.69
S34	Fish pop.							-0.10		-0.04		0.55	* 0.22
S33	Tissue							0.11		-0.01		0.53	* 0.40
S31	Pop. graph	0.22		0.55	*	0.00	0.49	-0.12		0.34		0.52	* 0.53
S19	Chem. change	0.16		0.09		0.52	* 0.56	0.28		0.00		0.50	* 0.52
S18	Storm	0.29		0.04		0.37	* 0.44	0.24		0.08		0.43	* 0.48
S22	Food chain	0.03		0.28	*	0.23	0.29	-0.08		0.29		0.43	* 0.43
S15	Respiration	-0.07		0.16		0.34	* 0.22	-0.08		0.08		0.36	* 0.14
S24	Model/obs.	0.18		0.15		0.28	* 0.33	0.12		0.22		0.31	* 0.36
S03	Chem graph	-0.12		0.49	*	0.30	0.45						
S20	Chem filter	-0.05		0.36		0.45	* 0.51						
S21	Ocean breeze	0.24		0.06		0.41	* 0.44						
S23	Chem react.	0.11		0.76	*	0.04	0.92						
S25	Guinea pig	0.18		0.34	*	0.08	0.32						
S32	Circuit	0.09		0.22	*	0.14	0.16						

Note. SM = Spatial-Mechanical Reasoning; QS = Quantitative Science; BKR = Basic Knowledge and Reasoning. * Indicates highest loading for each item.

Reasoning (BKR), consisting primarily of items that called for application of concepts and reasoning in biology and astronomy. Correlations among factors for 10th grade were 0.75 between SM and QS, 0.77 between SM and BKR, and 0.86 between QS and BKR. The 12th-grade correlations were 0.67 between SM and QS, 0.73 between SM and BKR, and 0.76 between QS and BKR. Correlations among corresponding EAP scores are, of course, lower; these are reported in a later section.

The results for the HSES sample are nearly identical to those obtained in the full NELS:88 sample (Hamilton et al., 1995; Nussbaum et al., 1997). The factor interpretations are based on inspection of item content and on observations of student responses obtained through interviews.

Distributions of Achievement

Means and standard deviations of scores on the science multiple-choice factors are given in Table 2 for males and females. The EAP scores derived from the full information item factor analysis are on a standard (mean 0, variance 1) scale. The table reveals that only SM shows a large gender difference, with males scoring nearly one half standard deviation higher than females. Gender differences on QS and BKR are minimal. It is worth reiterating that these results should not be interpreted as representative of a larger population (because sample weights are not used). They do indicate, however, that in this sample of students, SM exhibits substantial gender difference. This is consistent with findings reported in earlier work with the NELS:88 science tests (Hamilton et al., 1995).

Table 2

Descriptive Statistics by Gender on Multiple-Choice Factor Scores

	Females (N = 1080)				Males (N = 1090)			
	Mean	*SD*	Q1	Q3	Mean	*SD*	Q1	Q3
SM12	-.24	.984	-1.04	.56	.24	.960	-.52	1.09
QS12	-.01	.958	-.78	.69	.01	1.040	-.85	.88
BKR12	-.03	.992	-.82	.77	.03	1.008	-.61	.81

Note. SM = Spatial-Mechanical Reasoning; QS = Quantitative Science; BKR = Basic Knowledge and Reasoning.

Table 3 gives the frequencies of scores at each scale score level for each CR item, broken down by gender. The totals for each score reveal strong skewness, with relatively few students scoring at the highest levels. Especially noteworthy is the difference in numbers of males and females at score level 5 on Eclipses. Although more students achieved the highest possible score on this item than on the other three, the ratio of males to females is substantial. Similar but less extreme results are obtained for score levels 4 and 5 of Fuels and Populations.

Relations Among Scores on CR and MC Scales

Table 4 gives the Pearson product-moment correlations among the six MC achievement measures (three science factors, two math factors, and reading) at both 10th and 12th grades and the four CR scale scores. Table 4 reveals moderate correlations among all measures of achievement. Although the differences among coefficients are small, some patterns can be detected. Eclipses, for example, was more highly correlated with SM than with the other science factors, whereas Heating Curve had its highest correlation with QS. Reading achievement was more strongly related to performance on Fuels (CR1) than to the other three CR

Table 3

Frequencies of Constructed-Response (CR) Scale Scores by Gender

		Score						
Item		0	1	2	3	4	5	Total
CR1	Male	299	317	181	126	99	50	1072
	Female	435	340	131	82	55	25	1068
	Total	734	657	312	208	154	75	2140
CR2	Male	124	120	359	267	30	168	1068
	Female	229	234	357	174	19	50	1063
	Total	353	354	716	441	49	218	2131
CR3	Male	292	237	308	87	64	55	1043
	Female	366	266	256	81	47	37	1053
	Total	658	503	564	168	111	92	2096
CR4	Male	188	444	142	202	34	19	1029
	Female	180	414	195	208	25	15	1037
	Total	368	858	337	410	59	34	2066

Table 4

Correlations Among Constructed-Response (CR) Scale Scores and 10th- and 12th-Grade Multiple-Choice Factor Scores (N = 1551-2177)

	SM10	QS10	BKR10	MR10	MK10	RD10	SM12	QS12	BKR12	MR12	MK12	RD12	CR1	CR2	CR3	CR4
SM10	—															
QS10	0.64	—														
BKR10	0.61	0.60	—													
MR10	0.66	0.69	0.65	—												
MK10	0.55	0.61	0.60	0.76	—											
READ10	0.57	0.63	0.68	0.71	0.68	—										
SM12	0.69	0.52	0.55	0.60	0.50	0.49	—									
QS12	0.55	0.64	0.52	0.64	0.56	0.55	0.54	—								
BKR12	0.59	0.55	0.67	0.62	0.57	0.64	0.62	0.56	—							
MR12	0.40	0.40	0.42	0.54	0.50	0.46	0.51	0.45	0.58	—						
MK12	0.38	0.38	0.45	0.60	0.68	0.50	0.45	0.45	0.55	0.65	—					
READ12	0.53	0.58	0.63	0.64	0.62	0.79	0.54	0.55	0.70	0.58	0.56	—				
CR1	0.47	0.47	0.44	0.43	0.36	0.43	0.48	0.44	0.49	0.34	0.33	0.47	—			
CR2	0.46	0.38	0.38	0.38	0.31	0.31	0.47	0.37	0.40	0.32	0.27	0.33	0.39	—		
CR3	0.45	0.44	0.40	0.56	0.38	0.41	0.42	0.44	0.45	0.33	0.31	0.43	0.43	0.36	—	
CR4	0.39	0.41	0.39	0.40	0.38	0.39	0.39	0.43	0.41	0.30	0.33	0.42	0.39	0.29	0.39	—

Note. SM = Spatial-Mechanical Reasoning ; QS = Quantitative Science; BKR = Basic Knowledge and Reasoning; MR = Math Reasoning; MK = Math Knowledge; READ = Reading.

items, probably because Fuels involved an extended essay. In most cases, math reasoning (MR) was more highly correlated with science achievement than was math knowledge (MK). Despite these patterns, the correlations are difficult to interpret due to differences in reliability among the measures. Disattenuated coefficients are not presented because of the difficulty of accurately estimating reliabilities for the EAP scores.[3] Relationships among these measures were explored in greater detail through graphical procedures and multilevel modeling analyses; these results are not presented here.[4]

Differential Item Functioning: Multiple-Choice Test

Methods for identifying items exhibiting differential item functioning, or DIF, were applied to the 12th-grade MC and CR tests. This section reports results for the MC test. As is standard with tests of this nature, researchers at the Educational Testing Service (ETS) examined all items for DIF using the Mantel-Haenszel (MH) odds ratio procedure (Mantel & Haenszel, 1959). Instead relying solely on tests of statistical significance, ETS uses an "effect size" estimate (Zieky, 1993). A value "D" is defined as -2.35 times the log of the combined odds ratio across score levels. Items are labeled with "A" if D is not significantly different from zero or if the absolute value of D is less than 1. "B" items have D significantly different from zero and either absolute value of D less than 1.5 or absolute value of D not significantly different 1. "C" items, the only items for which DIF is considered practically important, have absolute value of D significantly greater than 1 and D larger than or equal to 1.5 (Camilli & Shepard, 1994). Rock and Pollack (1995) report only the number of "C" DIF items for each test at each grade. For gender-based DIF studies, only one science MC item at each grade (10th and 12th) exhibited "C" DIF.

Because the MH procedure treats the test as unidimensional, and because the earlier studies of the NELS:88 science test suggest the utility of treating it as multidimensional, it seemed worthwhile to explore other approaches. Therefore, the traditional MH method was applied and compared with a logistic regression procedure. The latter was conducted three times for each item: once

---

[3]The reliability of an EAP score varies by score level. Average reliabilities over all score levels could be calculated but there is a risk of overestimation, resulting overcorrection and spuriously high correlations.

[4]Multilevel modeling procedures were used to examine relationships among achievement, gender, and students' educational experiences. These are described in Hamilton (1997).

using total IRT score, once using the dimension from the full information item factor analysis on which the item loaded most highly, and once using all three dimensions. These analyses revealed ways in which definitions and interpretations of DIF changed depending upon the matching criteria used. This research uses the HSES sample rather than the full NELS:88 sample, on which the Psychometric Report is based; therefore some differences in results of the MH analysis should be expected.

Table 5 presents the results of the four DIF analyses for each item on the MC test. The first column gives the item numbers, the second provides brief descriptions, and the third lists the factor on which each item had its highest loading. P-values for males and females are provided in the fourth and fifth columns. These reveal variations in relative difficulty of items for males and females, with the largest differences occurring for SM items. The sixth column indicates which items showed statistically significant DIF using the MH procedure, with type I error rate of 0.01. The direction of DIF is indicated by "M" and "F." In addition, ETS-type effect sizes are given for each item.

The final three columns indicate statistically significant DIF using the logistic regression procedure with three sets of matching criteria. Instances of non-uniform DIF, which occurs when the magnitude of DIF varies by ability level, are indicated by "NU."

The MH procedure identified 14 items for which group members' probabilities of answering correctly differed significantly even when matched on total score. Half favored females and half males. Only one item had an effect size large enough to be categorized as "C;" five were categorized as "B." Not surprisingly, five of the seven items favoring males loaded on SM, and the only "C" item was the one with the highest loading on SM (lever). This item also had the largest difference in proportion correct for males and females: .72 versus .50. Three of the five "B" items loaded on SM. The remaining two items favoring males loaded on BKR. One of these was an astronomy item and the other dealt with indicators of an approaching storm. In contrast, of the seven items favoring females, five loaded on BKR, including one "B" item, and two on QS, also including one "B" item. The five BKR items involved content from biology as did one of the two QS items. The remaining QS item, with a "B" effect size, asked examinees to identify a product of radioactive decay. This is the only QS item that was easier on average for females than for males.

Table 5

DIF Results for Multiple-Choice Science Test

Item	Description	Dimension	*P*-value males	*P*-value females	MH	Effect size	IRT	Logistic regression	
								Single dim.	Three dim.
1	Simple reflex	BKR	0.90	0.89		A			
2	Moon's light	BKR	0.86	0.80	M	A	M	M	
3	Mix water	SM	0.77	0.69	M	A	M	F-NU	NU
4	Tissue	BKR	0.76	0.73		A			
5	Expt. design	BKR	0.78	0.81	F	B	F	F	
6	Earth orbit	SM	0.80	0.69	M	B	M		
7	Algae	BKR	0.70	0.64		A	M		
8	Classify subs.	BKR	0.70	0.67		A			
9	Storm	BKR	0.72	0.63	M	A	M	M	
10	Fish pop.	BKR	0.58	0.59	F	A		F	
11	Pop. graph	BKR	0.69	0.68	F	A			M
12	Contour map	SM	0.66	0.51	M	B	M		
13	Camera lens	SM	0.71	0.55	M	B	M		
14	Chem change	BKR	0.64	0.60		A			F
15	Respiration	BKR	0.59	0.63	F	A		F	
16	Food chain	BKR	0.51	0.53	F	A		F	
17	Uranium decay	QS	0.53	0.56	F	B		F	
18	Enzyme graph	QS	0.63	0.61	F	A			
19	Lever	SM	0.72	0.50	M	C	M	M	
20	Model/obs.	BKR	0.49	0.46		A			
21	Calc. mass	QS	0.43	0.37		A			
22	Half life	QS	0.46	0.39		A	M		
23	Pendulum	SM	0.45	0.40		A			
24	Hydro. react.	QS	0.36	0.34		A			
25	Train	SM	0.17	0.16		A		NU	NU

Note. SM = Spatial-Mechanical Reasoning ; QS = Quantitative Science; BKR = Basic Knowledge and Reasoning; M indicates DIF in favor of males. $p < .01$. F indicates DIF in favor of females, $p < .01$. NU indicates non-uniform DIF, $p < .01$.

These results were compared with those obtained when logistic regression was conducted using science IRT score as matching criterion. Eight of the fourteen items identified by MH were also flagged by the logistic procedure; seven of these favored males, including five SM items. Two additional items were flagged as favoring males. In contrast to the MH procedure, nine out of ten flagged items favored males. No cases of non-uniform DIF were observed. Again, the most serious DIF is associated with the SM items. Simulation studies have shown that MH and logistic regression function similarly (Swaminathan & Rogers, 1990); therefore, the difference observed here is probably due to the different conditioning variables (number-correct score versus IRT score). These two analyses produced similar results, however, in that both revealed that SM items exhibited the largest DIF.

The third analysis conditioned on a single dimension from the full information factor analysis. Five of the ten items identified with IRT matching were flagged. One of these, the item dealing with mixing water of different temperatures, favored females instead of males under this method and showed significant non-uniform DIF. Four additional items showed DIF in favor of females. One item (25), which showed no DIF in either of the previous analyses, exhibited significant non-uniform DIF. This item appeared to combine spatial ability and physics knowledge and had large loadings on both SM and QS. Matching on SM eliminated the male advantage on four items, but item 19 (lever) still favored males. This method is flawed because a single dimension does not capture the variety of abilities measured by a particular item, especially for items with high loadings on more than one dimension. It is informative, however, to discover how the perception of DIF changes with the definition of ability.

Finally, all three dimensions were included as conditioning variables, along with their interactions with gender. This procedure identified the fewest items; two were flagged as exhibiting uniform DIF, one favoring females and one males. The item favoring females asked examinees to identify a process that represented a chemical change, and the item that favored males involved interpretation of growth rates from a graph. Two additional items showed non-uniform DIF. One of these was a physics item that asked students to identify the path of a ball dropped in a moving train. It had fairly high loadings for all three dimensions. All three interaction terms were positive, indicating that the

relationship between performance on each factor and score on this item was stronger for males than for females. The other item asked students to select the temperature of a water mixture. The three negative interaction terms on this item reveal a weaker relationship between performance on each factor and score on the item for males than for females, especially for SM. In other words, the probability that a female will select the correct response on this item is more heavily dependent on her overall spatial-mechanical ability (as measured by this test) than is the probability of a correct response for a male. This relationship is masked when DIF is investigated by conditioning on total score alone. The finding that fewer items were identified as exhibiting DIF when all three MC factors were included as matching criteria reveals the importance of conditioning on all abilities measured by a test; items identified as exhibiting DIF in the previous analyses were, in most cases, measuring one or more specific abilities that were not captured by total score or by a single dimension.

Differential Item Functioning: Constructed-Response Test

DIF detection on the CR test presents two major challenges: (1) the absence of a reliable matching criterion in the form of total score on a test of similar items, and (2) the polytomous scoring system. The second problem can be addressed by the use of the logistic discriminant function analysis (LDFA) procedure (Miller & Spray, 1993). This study investigates solutions to the first problem by examining the effects of including various sets of conditioning variables.

The LDFA procedure was carried out several times for scale scores on each of the four CR items, using various combinations of science achievement measures as matching criteria. CR total score (which included the studied item) and MC IRT score were each investigated separately. Because BKR appears to be the most similar to general science achievement of the three MC dimensions, it was included alone as a matching variable. Then QS was added, followed by SM, which is the least like a general science achievement measure. The MC scores and CR total score were also entered in combination. Table 6 indicates for which items and matching criteria significant DIF was revealed. For Fuels and Populations, the only factor determining whether or not DIF was present was the inclusion of CR total score as a matching criterion. When CR total score was excluded, the LDFA procedure suggested the presence of DIF in favor of males for both items. Eclipses and Heating Curve showed DIF regardless of matching

Table 6
DIF Results for Constructed-Response (CR) Science Test

Matching Criteria	CR1	CR2	CR3	CR4
CR Total		M		F
IRT	M	M	M	F
BKR	M	M	M	F
BKR+QS	M	M	M	F-NU
BKR+QS+SM	M	M	M	F
CR+IRT		M		F
CR+BKR		M		F
CR+BKR+QS		M		F
CR+BKR+QS+SM		M		F

Note. BKR = Basic Knowledge and Reasoning; QS = Quantitative Science; SM = Spatial-Mechanical Reasoning ; M indicates DIF in favor of males, $p < .01$. F indicates DIF in favor of females, $p < .01$. NU indicates non-uniform DIF, $p < .01$.

criteria, with Eclipses favoring males and Heating Curve favoring females. The only instance of non-uniform DIF occurred for Heating Curve when BKR and QS were included as matching criteria. It appears, therefore, that Eclipses and Heating Curve warrant further investigation.

To assess the importance of the DIF on these two items, Scheffe-type confidence bands were constructed around the logistic discriminant function curve at each score level of each of these two items. For simplicity, only one conditioning variable is included in each set of plots; confidence bands were constructed for models that included only CR total score as conditioning variable and also for those that included only MC IRT score. This procedure is somewhat conservative (Hauck, 1983), so a type I error rate of 0.05 was used for these analyses. Each plot shows the null model (probability of being male given CR total score) and the full model (probability of being male given CR total score and item score, with interaction term included). Confidence bands are given for the full model; score regions where the curve for the null model lies outside the confidence bands indicate practically important DIF (Miller & Spray, 1993). Figure 1 shows 95 percent confidence bands for each score level of Eclipses for the total CR score model, and Figure 2 gives confidence bands for Eclipses when MC IRT score is used. Figures 3 and 4 show similar plots for Heating Curve. For

illustrative purposes, all six score levels are plotted for Figure 1, but the remaining figures show plots only for the lowest and highest score levels.

The curve for the null model in Figure 1 shows that the predicted probability of being male is strongly related to total CR score. When Eclipses score is included in the model, however, the relationship between probability of being male and total CR score is greatly diminished. Throughout most of the CR total score range, students receiving low scores on Eclipses were significantly less likely to be classified as male than they would have been if Eclipses score were ignored; students receiving high scores on Eclipses were more likely to be classified as male. It should be noted that because of the dependency between Eclipses and CR total, some high and low scores in these plots are impossible; for example, students scoring 3 on Eclipses cannot receive a CR total score higher than 18.

The finding of DIF on this item may result, in part, from the mutual influence of Eclipses and Heating Curve, which show DIF in opposite directions (Wang & Lane, 1996). Therefore, it is worthwhile to use a matching criterion that does not include these two items. Using CR score is not feasible because only two items would remain. Instead, MC IRT score is used alone as matching criterion to examine the practical importance of DIF on these items. This method is less than ideal because MC score cannot be interpreted as representing the ability measured by the CR test. Nonetheless, it is informative to show the relationship between gender and CR item scores for students matched on MC score. Confidence bands for the logistic discriminant function curve for each level of Eclipses score are given in Figure 2.[5]

These plots reveal a weaker relationship between Eclipses and MC than between Eclipses and total CR score. This is to be expected because of the format differences but also because of the dependency of CR total on Eclipses. Still, students receiving low scores on Eclipses are more likely to be classified as female than they would be based only on their MC scores; this is especially true for those who receive high MC scores. At higher score levels on Eclipses, students receiving low MC scores are more likely to be classified as male than they would be if Eclipses score were ignored.

---

[5]The range of IRT scores is from 10 to 35. IRT score provides an estimate of the total number of items the student would have answered correctly if he or she had taken all of the items that appeared on any version of the science test at any grade.

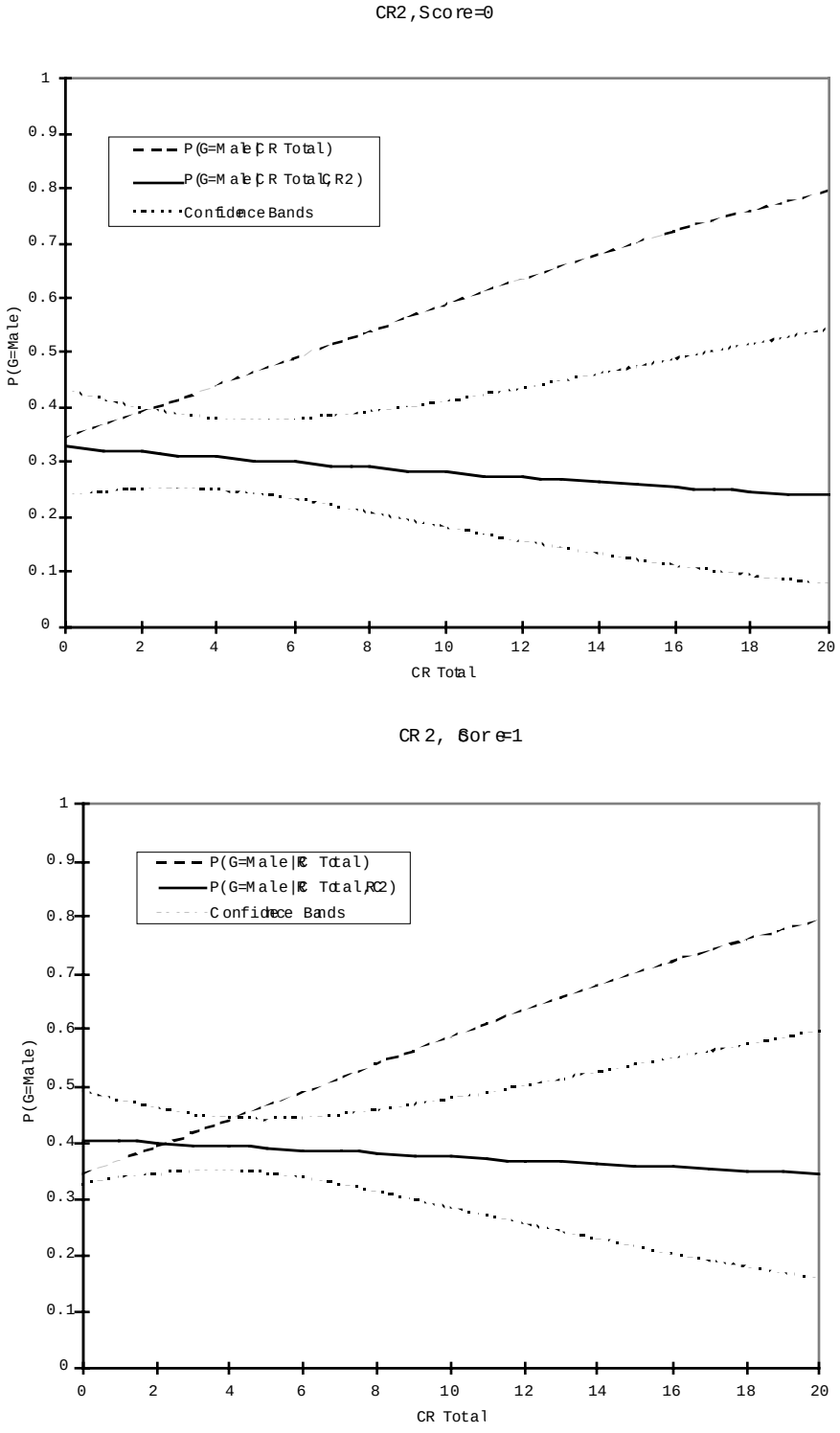
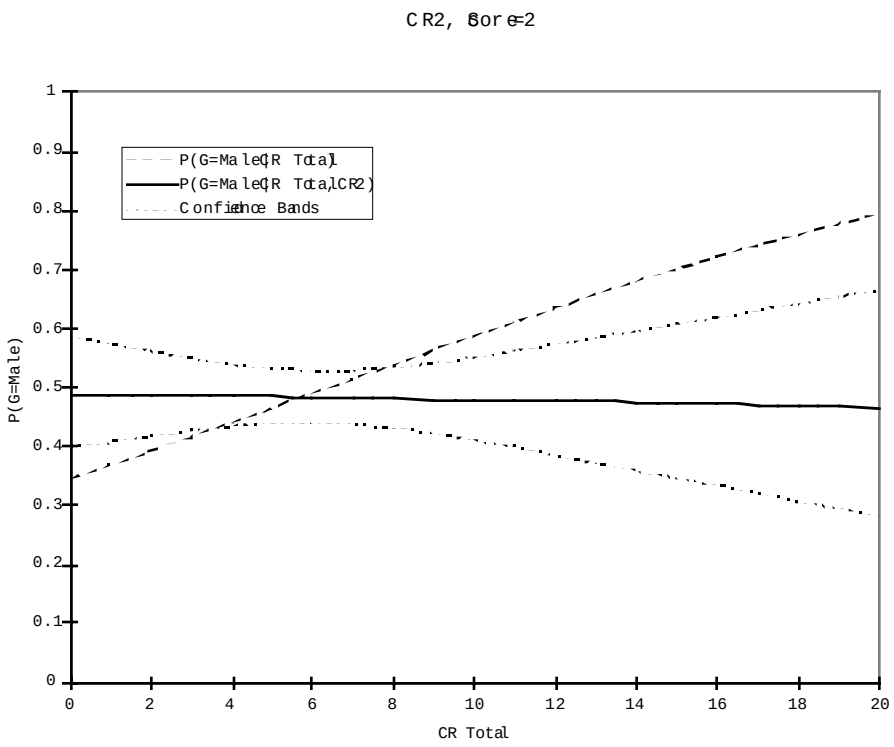


Figure 1. 95% confidence bands for LDFA curve, CR Item 2 (Eclipses), using CR total score as matching criterion.



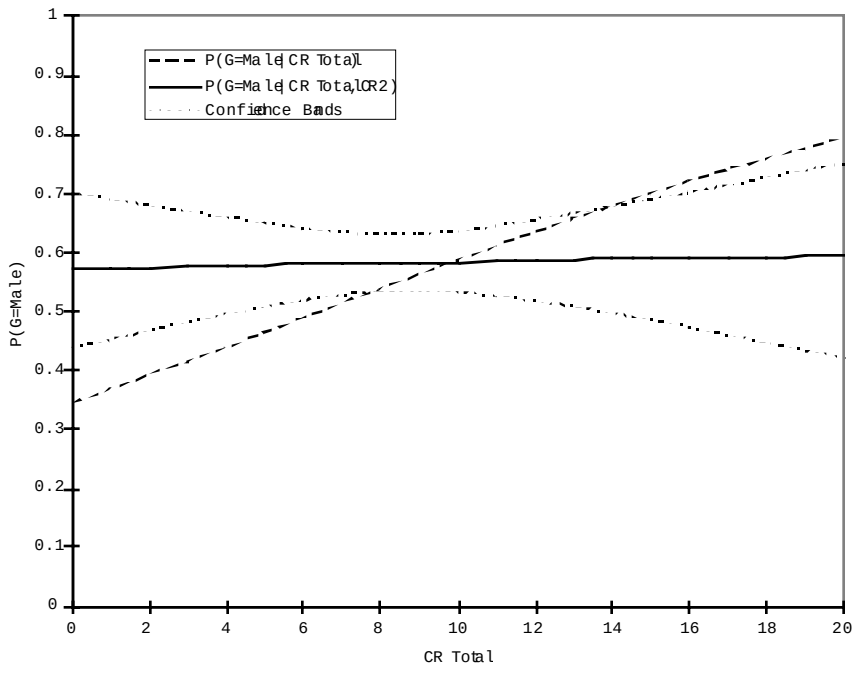
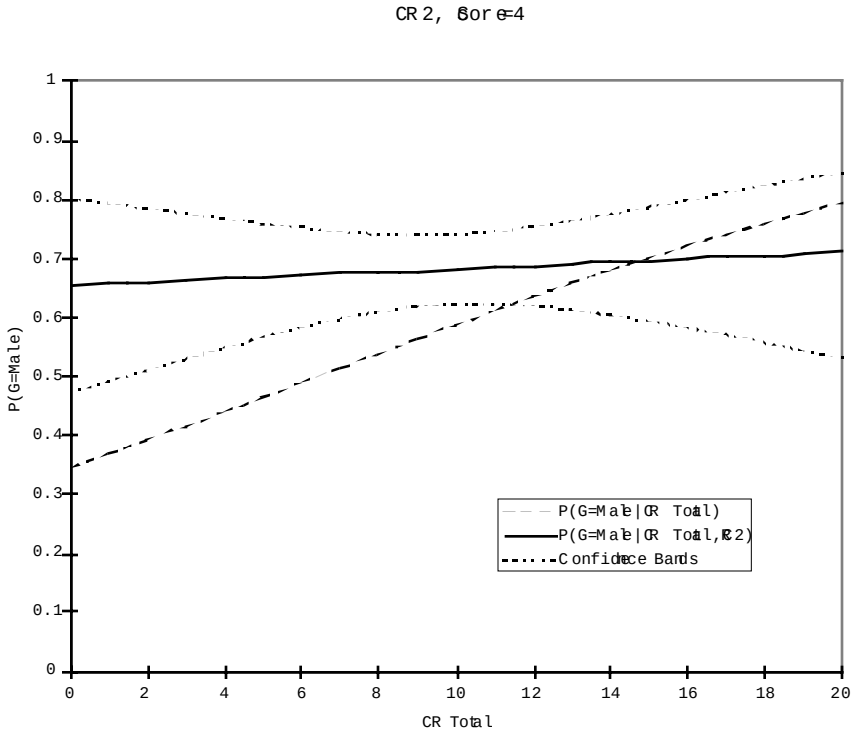


Figure 1. (continued).



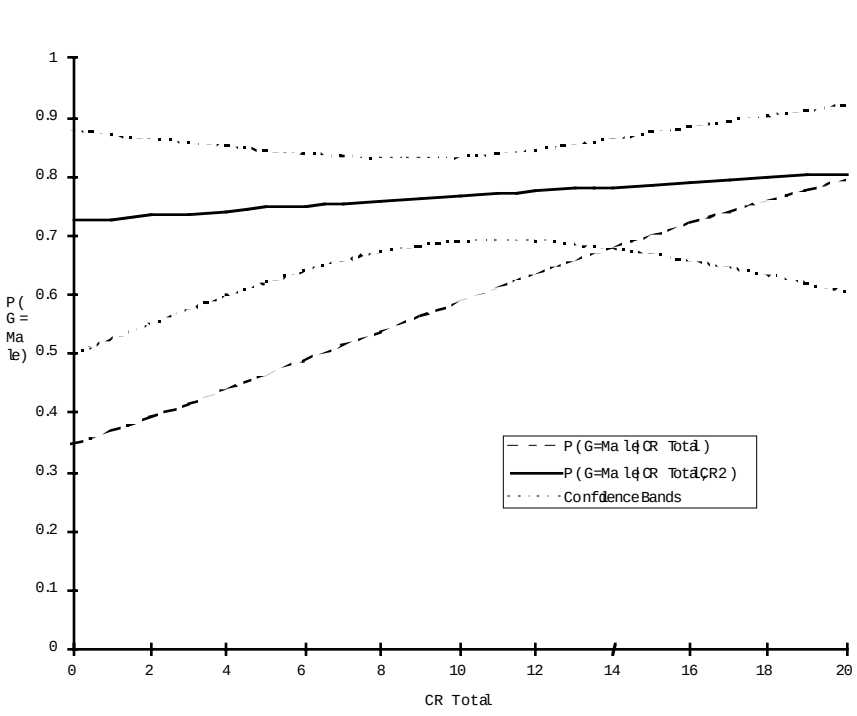


Figure 1. (continued).

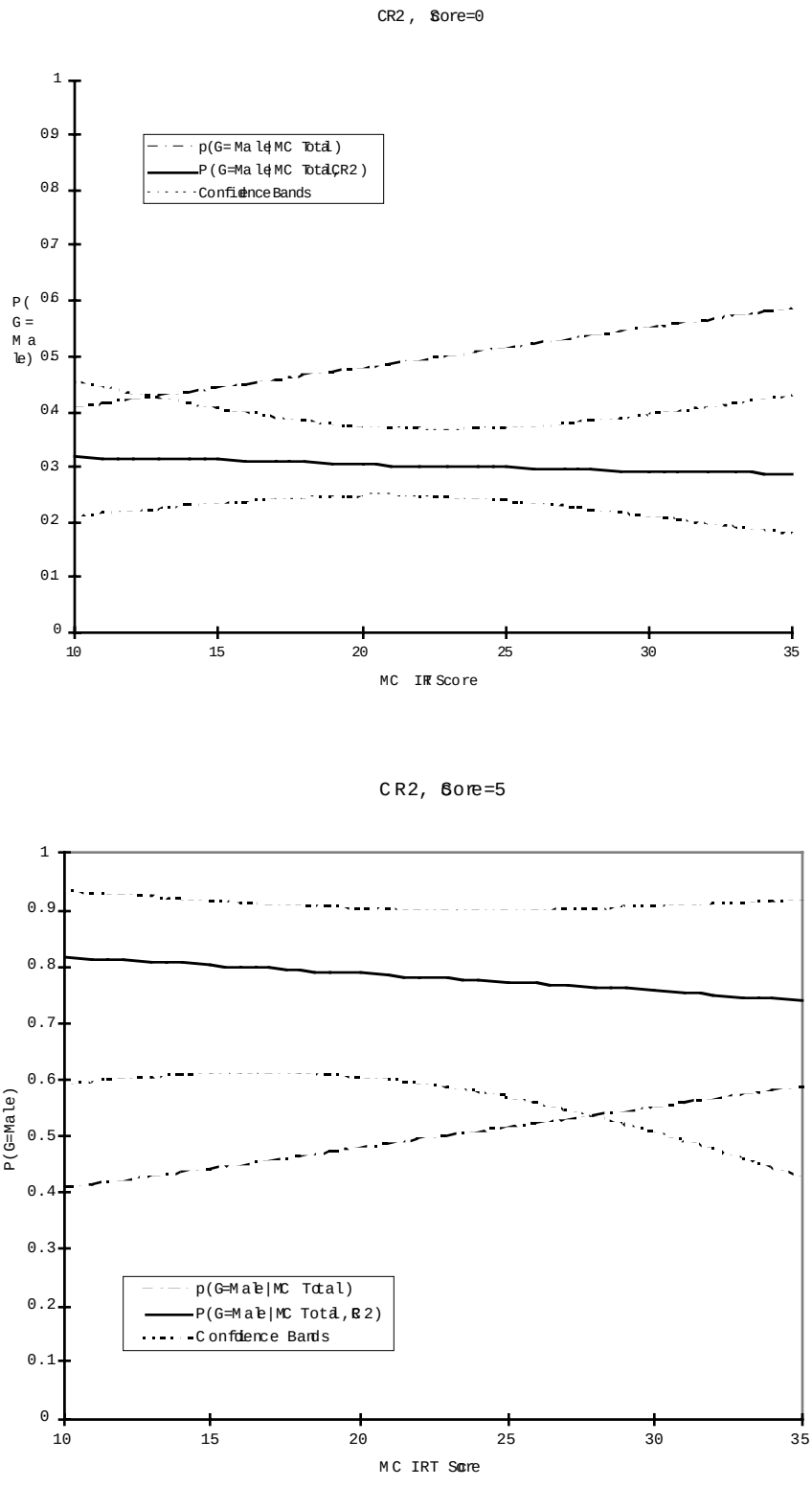


Figure 2. 95% confidence bands for LDFA curve, CR Item 2 (Eclipses), using MC IRT score as matching criterion.

Confidence bands for Heating Curve based on total CR score appear in Figure 3. The pattern here is opposite that for Eclipses. The curve for the null model lies below the lower confidence band at high CR total score levels and low Heating Curve levels, and above the upper confidence band at low CR total levels and high Heating Curve levels.

Unlike Eclipses, however, the results based on MC IRT score do not appear to be practically significant. Figure 4 shows that at each level of Heating Curve, the function for the null model lies within the 95 percent confidence bands for the full model. In other words, knowing one's score on Heating Curve apparently provides no additional information about one's probability of being male once total MC score is known. Nonetheless, the previous finding of DIF favoring females on this item suggests that sources of gender differences should be investigated.

Cumulative Logits Analysis

Gender differences on CR scale scores may be larger at some score level transitions than at others; for example, perhaps males and females are equally likely to reach some minimum level of performance but more males than females achieve scores of 4 and 5. Furthermore, even items that do not exhibit DIF when scale score is studied may show important differences at particular score levels. To explore these possibilities, DIF analyses were conducted for each of five possible scale score splits on each CR item. This procedure is known as cumulative logits analysis (Agresti, 1990; French & Miller, 1996). Scores were divided into two groups for each analysis: those scoring at or below a certain level and those scoring above. Three separate sets of conditioning variables were used for each analysis: CR total score, CR total plus MC IRT, and CR total plus three MC dimensions. The analyses were conducted using logistic regression, and both uniform and non-uniform DIF were investigated.

Consistent with the findings from the previous analysis, neither Fuels nor Populations showed DIF at any score level[6]. Eclipses showed DIF in favor of males, regardless of matching criteria, at every score level split except that between 0 and 1. In other words, equally able (as measured by the matching variables) males and females do not differ in their probabilities of attempting the problem and providing at least an incomplete explanation on Part C (the

[6]Tables are not presented for these analyses but are available from the author.

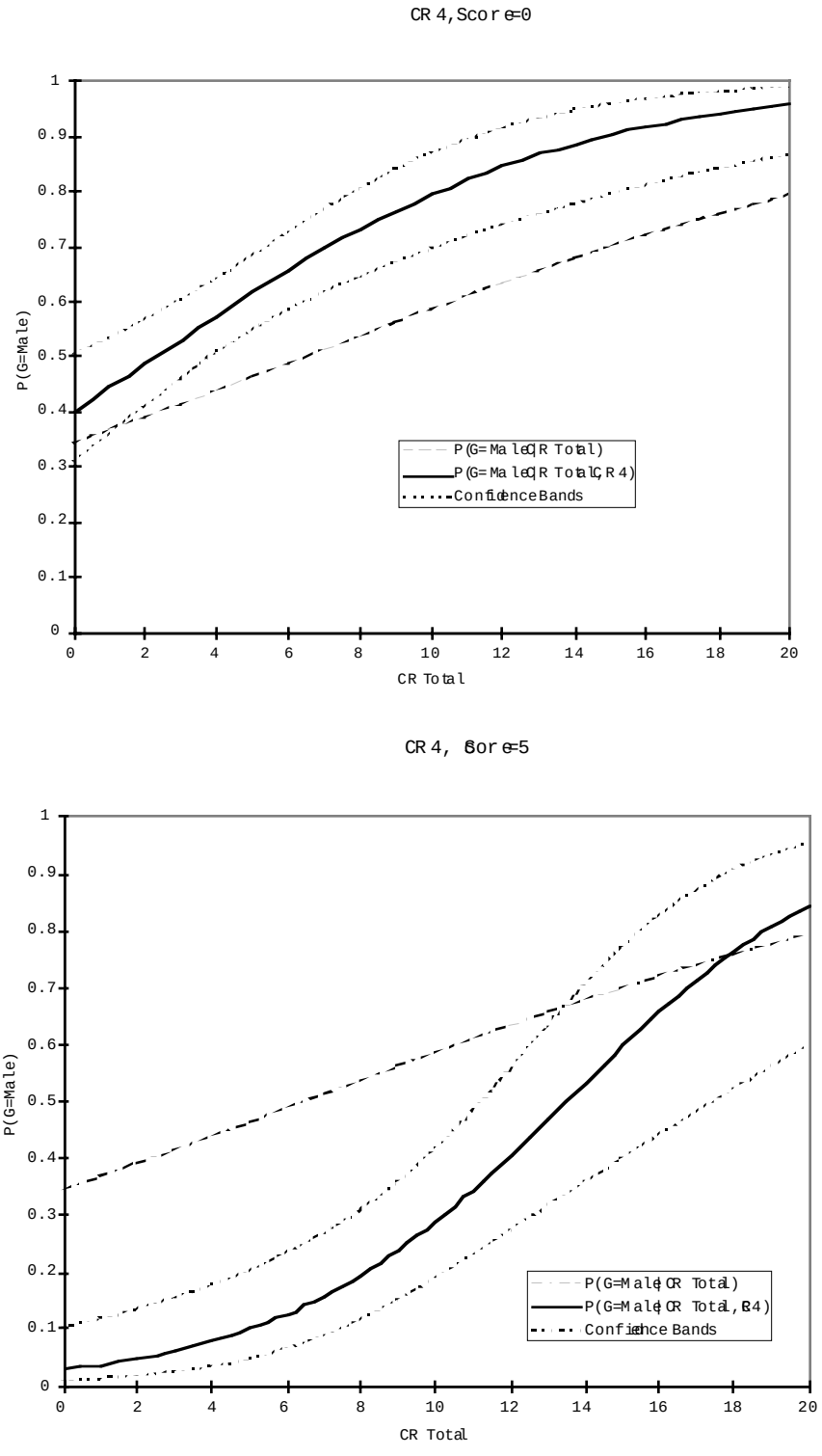


Figure 3. 95% confidence bands for LDFA curve, CR Item 4 (Heating Curve), using CR total score as matching criterion.

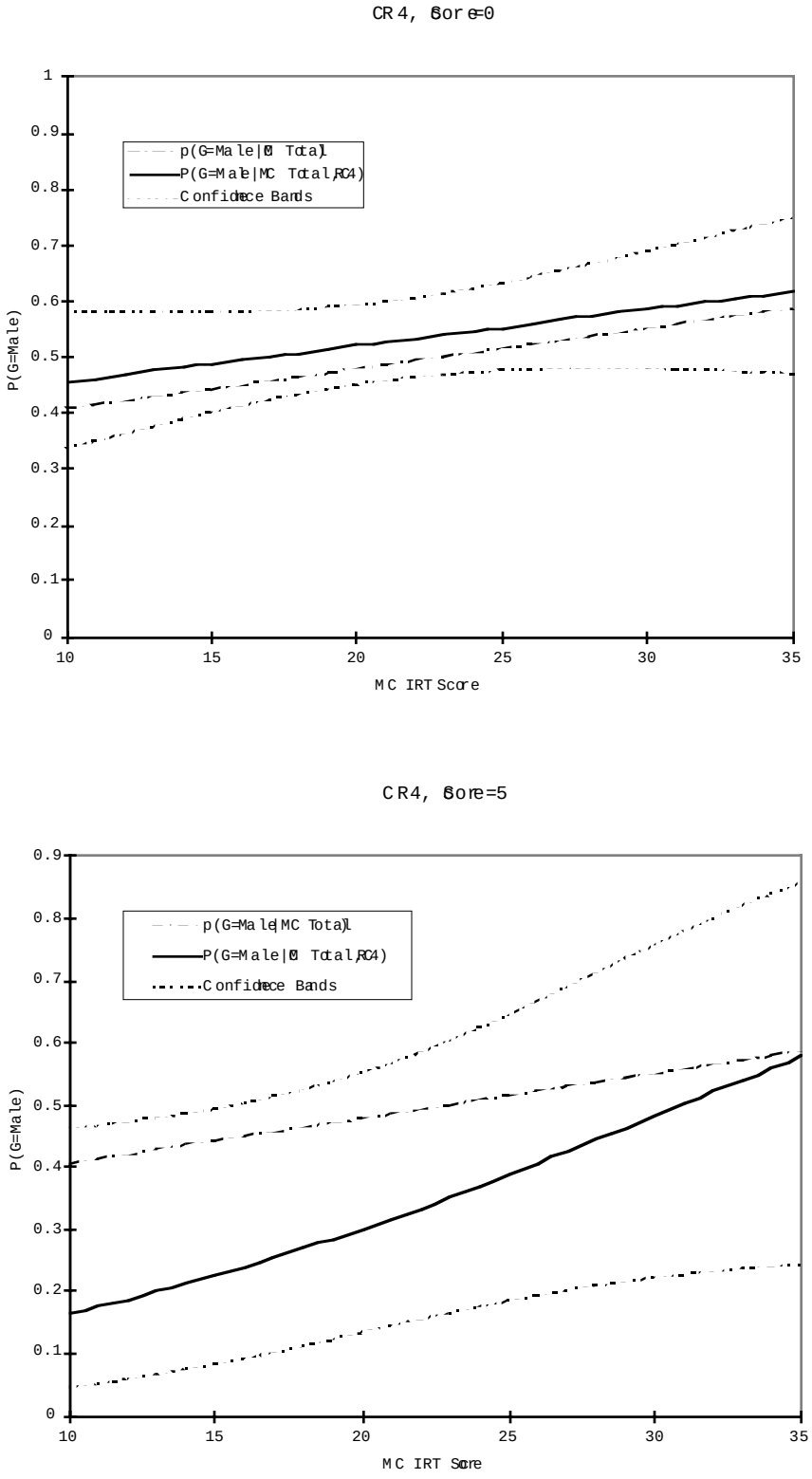


Figure 4. 95% confidence bands for LDFA curve, CR Item 4 (Heating Curve), using MC IRT score as matching criterion.

requirements for a score of 1), but they do differ at all other score levels. Non-uniform DIF was not observed. Heating Curve, which showed DIF favoring females when scale score was used, favored females at the lowest three splits but not at the highest two. This suggests that females are more likely than their equally able male counterparts to provide at least a partially correct response but are no more likely to achieve the highest scores. The next section provides more detail on these findings by analyzing categorical subscores separately.

Categorical CR Scores

As discussed earlier, scale scores may mask important information contained in responses to specific parts of an item; the analytical subscores provide some data about these responses. The analytical scores are categorical in nature, but many could be easily dichotomized (e.g., categories of responses to the solar eclipse diagram could be divided into two groups: drew correct diagram or not). DIF analyses for each categorical score were conducted using the conditioning variables described in the previous section. The results were not affected by the choice of conditioning variables. There were two main findings. First, on Eclipses, significant DIF in favor of males was observed for all three parts of the item: drawing the solar eclipse, drawing the lunar eclipse, and writing the explanation. In other words, males were more likely to produce accurate diagrams of both the solar and lunar eclipse than were females, and were more likely to provide a correct explanation. Males and females were equally likely, however, to provide at least a partial explanation. The other main result was for Heating Curve, in which DIF in favor of females was observed for only the first part of the item (describing what was happening as the ice began to melt). However, in contrast to the results for Eclipses, females were more likely than males to supply a response to this item that was at least partially correct, but were no more likely to provide a completely correct response. Significant DIF was not observed on the other parts of the item. These results suggest that the relative weighting assigned to various parts of the items could affect the extent to which DIF is observed, but the effects are not likely to be great. Eclipses showed the strongest DIF of the four CR items, and all three of its parts exhibited DIF.

Small-Scale Interview Study

The DIF detection procedures reveal items that exhibit unusually large gender differences, and inspection of these items can suggest features that

account for these differences. However, it is difficult through inspection alone to determine what the items are measuring and what processes and sources of knowledge are evoked. The interviews supply some of these details. The small sample used in the interview study precludes analysis of differences in strategies among males and females; instead the objective was to identify salient features of the items that showed the largest difference in scores in the national sample.

Multiple-Choice Test

Of the 16 MC items included in the interview study, seven loaded on SM, four on QS, and five on BKR. Table 7 presents frequencies of coding categories across the three dimensions. The raw frequency for a category is simply the number of times that response was observed across all items on the factor and across students. Corresponding probabilities were calculated by dividing the raw frequency by 25 (the number of students in the sample) and by the number of items on the factor. These values therefore reflect the probability that a student will exhibit a particular response to an item loading on a particular factor. For example, the numbers in the first row of Table 7 show that "quick elimination" was used four times on the SM items, and the corresponding probability is 4/(25*7) = .02. Items loading on the SM dimension, which showed the largest gender differences in the national sample, could be distinguished from BKR and QS items primarily by students' use of prediction, gestures, and visualization. A response was coded as involving prediction when a student discussed what would happen under various conditions. The pendulum item, for example, was frequently associated with statements such as, "If I lengthen the string, it will swing more slowly." Several of the SM items involved this type of reasoning. Reports of visualization were not as prevalent as the use of gestures and prediction, but visualization clearly played a larger role in SM items than in others. Visualization was often observed in conjunction with prediction, as when a student reported a mental picture of the effects of various manipulations. The use of gestures is consistent with other research demonstrating that gestures are more likely to be observed when speech contains spatial content than when it does not (Rauscher, Krauss, & Chen, 1996). Physics was the course most strongly associated with these items. Students were also more likely to report using information from laboratory and hands-on activities, as well as outside-of-school experiences such as hiking and reading maps.

Table 7

Frequencies and Probabilities of Responses to Multiple-Choice Items by Dimension

Coding category	Raw frequency			Probability		
	SM	QS	BKR	SM	QS	BKR
Quick elim.	4	2	9	.02	.02	.07
Consider alt.	34	37	48	.19	.37	.38
Calculation / graph	21	60	24	.12	.60	.19
Scientific expl.	13	12	32	.07	.12	.26
Made sense	61	9	53	.35	.09	.42
Prediction	72	7	0	.41	.07	.00
Gestures	65	1	2	.37	.01	.02
Visualization	31	1	3	.18	.01	.02
Guess	7	15	5	.04	.15	.04
Guess part	3	3	9	.02	.03	.07
Vocab. problem	8	13	13	.05	.13	.10
Not sure	8	15	6	.05	.15	.05
Biology	7	18	43	.04	.18	.34
Chemistry	13	51	13	.07	.51	.10
Physics	48	7	7	.27	.07	.06
Earth science	4	2	2	.02	.02	.02
Elementary sci.	25	3	23	.14	.03	.18
Math course	21	11	11	.12	.11	.09
Hands-on or lab	43	0	6	.25	.00	.05
Book	2	1	2	.01	.01	.02
Outside school	19	0	11	.11	.00	.09
Never learned / not sure where learned	51	31	35	.29	.31	.28

Note. SM = Spatial-Mechanical Reasoning; QS = Quantitative Science; BKR = Basic Knowledge and Reasoning.

"Playing with" equipment such as see-saws or cameras seemed to be an especially powerful kind of learning activity.

Coding category frequencies were studied separately for successful and unsuccessful students, but both groups displayed approximately equivalent use of visualization, gestures, and prediction. It appears that SM items tend to evoke particular kinds of responses but that the quality of these activities varies. The interviews revealed that the SM items depend heavily on visual or spatial reasoning combined with knowledge acquired in school or through

extracurricular activities, and that hands-on activities may be especially beneficial. The male advantage on these items may result from the spatial reasoning demands and, in particular, from differences in exposure to activities that promote visual or spatial reasoning.

Constructed-Response Test

The item of primary interest with regard to the DIF study is Eclipses. The major difference between Eclipses and the other CR items was that most responses to Eclipses included evidence of visualization. Only four students reported no visualization, and their average score was 0.25, compared with a mean of 3.86 for the students who did report visualization. The non-visualizers said they had no idea how to approach the problem or had never learned about eclipses, so they probably lacked the knowledge needed to form a visual image. Successful students reported forming mental images of the solar system; e.g., "I just thought about it and imagined what it looks like." Many students also used gestures when describing their reasoning, and gestures are often observed in conjunction with spatial material (Rauscher, Krauss, & Chen, 1996). These responses suggest that Eclipses elicited a form of spatial reasoning similar to that used on the spatial-mechanical reasoning (SM) multiple-choice dimension, which was the only dimension on the MC test to exhibit a large gender difference. Successful responses to Eclipses typically involved a combination of knowledge acquired in or outside of school with reasoning of a visual or spatial nature. The gender difference on Eclipses is therefore not surprising, given the large gender differences on most of the SM items. Of course SM and Eclipses did not function identically, and Eclipses showed DIF even for students who were matched on SM. The spatial demands may have been greater on Eclipses than on similar MC items because of the unstructured nature of the CR items.

Most students reported using knowledge learned outside of school to complete this item. This was particularly true for the lunar eclipse diagram; 13 of the 16 students who produced a correct diagram said they had never learned about lunar eclipses in school. These students reasoned from other parts of the item (e.g., "I know in a solar eclipse the moon is between the sun and the earth, so a lunar eclipse must be the other way around and have the earth between the moon and the sun") or relied on information to which they had been exposed through television, books, newspapers, or actual eclipse viewing. Again, this

item is similar to SM MC items, which also elicited reports of using outside experiences.

After Eclipses, Fuels was the item that showed the largest gender difference in its raw distribution. Interestingly, this item favored males even though it consisted solely of a single essay and thus might have been expected to favor females. However, analysis of the interview results as well as examination of the scoring rubric revealed that writing ability had little or no effect on scores. Students received points for mentioning at least one advantage and one disadvantage of each type of fuel, and were not rewarded for organization, mechanics, or style. Interview respondents rarely displayed evidence of the kind of planning that usually accompanies writing tasks: 17 of the 25 started writing immediately after reading the question, and even those who did some planning tended to list the major points quickly and then write them down.

As on Eclipses, students reported using outside knowledge more often on this item than on either of the remaining CR items (Heating Curve and Populations). Thus one consistent feature of the items favoring males in either format is an apparent need to apply knowledge or skills beyond those currently emphasized in most science classrooms. Further evidence for this conjecture is obtained when performance on the nuclear and fossil fuels parts of the item are studied separately: The gender difference on this item was due to males' superior performance on the discussion of nuclear fuels; no difference occurred for fossil fuels. While most students said they had studied fossil fuels in school, knowledge about nuclear fuels was typically obtained through outside reading or television viewing. The difference between scores of males and females on this item, then, arose solely from differences on the part of the item that tended to call on outside knowledge. This result suggests that efforts to improve the science achievement of females might profitably consider ways in which such extracurricular experiences could be incorporated into formal science instruction.

Heating Curve was the only item that exhibited DIF in favor of females. Its raw distribution showed virtually no gender difference, in contrast to the other three CR items, which all favored males. The interviews revealed that the distinguishing feature of this item was its similarity to an activity that students had encountered in class. Eighty percent of participants said they had conducted an experiment similar to the one described in this item, in sharp contrast to the other CR items, which were unfamiliar to most students. Also in contrast to the

other items, knowledge acquired outside of school did not appear to enhance performance on this item. In fact, both male and female students who referred to outside experiences, such as boiling water in the kitchen, tended to receive low scores because their written responses did not include the kind of scientific terminology (such as reference to potential and kinetic energy) that high scores required. These results are consistent with other research that has demonstrated relative female advantage on items that resemble textbook material or that are closely tied to curriculum (Hanna, 1989; O'Neill & McPeek, 1993). The Heating Curve results provide additional support for the assertion that the items most likely to favor males are those that are the least closely tied to school curriculum.

Summary

The interviews provided support for the hypothesis that the SM items and Eclipse had in common some dependence on visual or spatial reasoning. The interviews also revealed the importance of knowledge acquired outside of school or through hands-on activities. This information is important for test developers or users who seek to understand the nature of gender differences in measured science achievement.

Discussion

In this study, gender differences on two kinds of science achievement measures were discussed. In some cases these differences were large enough to be considered important; most of these favored males over females. The size of the gender difference varied across subscales of the MC test and across items within the CR test; some types of items were more susceptible to gender differences than were others.

The DIF analysis revealed particular items within the tests that showed especially large differences, suggesting that these items measured some construct that is related to gender. Varying the matching criteria affected the identification of DIF on the MC test. In particular, taking into account the multiple constructs measured by the test reduced the number of items flagged as exhibiting DIF. Items with spatial or visual content showed the strongest effects when total score was used alone, but conditioning on spatial-mechanical reasoning eliminated the DIF. Users of multiple-choice tests should be aware of the multidimensional nature of many such tests when interpreting performance differences between

population subgroups. In the case of NELS:88, the significant male advantage on total score is due to performance differences on one type of item and not to overall superiority in science.

Choice of conditioning variable also influenced DIF detection on the CR test to some degree. However, the DIF identified on two CR items persisted regardless of matching criteria. Better measures of the abilities measured by CR tests such as this one are needed to explore DIF adequately. The results reported here, however, do provide some important information for users of large-scale assessment results. The CR format does not necessarily reduce gender differences in science achievement and may, in fact, increase them under certain circumstances. For both the MC and CR tests, perhaps the central message should be that using total test score masks differences among items within a test. Conclusions concerning relationships between achievement and group membership or educational background are influenced by the achievement measure used. For some purposes, it may be appropriate to consider items or subsets of items to acquire more detailed information about such relationships.

This study raises questions concerning the sources of gender differences on particular items. Is the male advantage on SM items and on the Eclipse problem due to differences in course-taking or exposure to science outside of school, or can it be attributed to a more highly developed spatial ability that is formed in the elementary school years or even earlier? Additional research is needed, but the combination of statistical and interview analyses reported here provides some hypotheses concerning sources of gender differences and ways to reduce them.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching, 26,* 141-169.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88,* 365-377.

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

Bock, D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal, 34,* 297-331.

Camilli, G., & Shepard, L. A. (1994). *Methods for detecting biased test items.* Thousand Oaks, CA: Sage Publications.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202-214.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1997). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33,* 453-464.

Cole, N. S. (1997, May). *The ETS gender study: How males and females perform in educational settings.* Princeton, NJ: Educational Testing Service.

Douglas, J. A., Roussos, L. A., & Stout, W. (1997). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33,* 465-484.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-303.

Fennema, E., & Tartre, L. A. (1985). The use of spatial visualization in mathematics by girls and boys. *Journal for Research in Mathematics Education, 16,* 184-206.

Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching, 20,* 481-495.

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

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.

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52,* 1091-1102.

Hamilton, L. S. (1997). *Construct validity of constructed-response assessments: Male and female high school science performance*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181-200.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. E. (1995). Enhancing the validity and usefulness of large scale educational assessments: II. NELS:88 science achievement. *American Educational Research Journal, 32,* 555-581.

Hanna, G. (1989). Mathematics achievement of girls and boys in grade eight: Results from twenty countries. *Educational Studies in Mathematics, 20,* 225-232.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. *The American Statistician, 37,* 158-160.

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. S., 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 Service.

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,* 352-366.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal, 32,* 525-554.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large scale educational assessments: III. Mathematics performance through the twelfth grade. *American Educational Research Journal, 34,* 124-150.

Linn, M. C. (1985). Fostering equitable consequences from computer learning environments. *Sex Roles, 13,* 229-240.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17-19, 22-27.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1-10). New York: Macmillan.

Lohman, D. F. (1993). *Spatially gifted, verbally inconvenienced*. Invited address to the Wallace Symposium on Talent Development, University of Iowa.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131-144.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations* (College Board Report No. 92-7). New York: College Entrance Examination Board.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: Macmillan.

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.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1991). *The state of mathematics achievement: Executive summary* (Report 21-ST-04). Washington, DC: National Center for Education Statistics.

National Assessment of Educational Progress. (1988). *The science report card: Elements of risk and recovery. Trends and achievement levels based on the 1986 National Assessment.* Princeton, NJ: Educational Testing Service.

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

National Council on Education Standards and Testing. (1992). *Raising standards for American education*. Washington, DC: U.S. Government Printing Office.

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large scale educational assessments: IV. NELS:88 Science performance through the twelfth grade. *American Educational Research Journal, 34,* 151-173.

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

Pollack, J., & Rock, D. (1995). *Constructed-response items in the NELS:88 High School Effectiveness Study*. Washington, DC: National Center for Education Statistics.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7,* 226-231.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 35-75). Boston: Kluwer.

Rock, D. A., & Pollack, J. M. (1995). *Psychometric report for the NELS:88 base year through second follow-up* (NCES Report No. 95-382). Washington, DC: National Center for Education Statistics.

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.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40,* 106-108.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9,* 175-199.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32,* 163-178.

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? *Journal of Research in Science Teaching, 31,* 857-871.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

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