Identifying Differential Item Functioning on the NELS:88 History Achievement Test

CSE Technical Report 511

Vi-Nhuan Le
CRESST/Stanford University

October 1999

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—Psychometric and Cognitive Modeling Richard J. Shavelson, Project Director, CRESST/Stanford University

The work reported herein was supported 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 the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

IDENTIFYING DIFFERENTIAL ITEM FUNCTIONING ON THE NELS:88 HISTORY ACHIEVEMENT TEST

Vi-Nhuan Le
CRESST/Stanford University

Abstract

This study examined gender-based differential item functioning (DIF) on the 10th-grade history achievement test administered as part of the National Education Longitudinal Study of 1988 (NELS:88). Several DIF analyses with varying matching criteria were conducted, and results were supplemented with a survey study that helped validate the interpretations of the underlying causes of DIF. DIF in favor of each gender corresponded to traditional sex-role stereotypes; males performed better on "masculine" items, whereas females were advantaged on "feminine" questions. The survey study confirmed that both high school boys and high school girls perceived the items to be sex-typed in the manner predicted by sex-role appropriateness. The findings revealed that the male advantage on this particular test was limited to specific content areas and did not represent a difference in overall proficiency.

In the past few decades, considerable attention has been paid to gender differences on standardized tests of achievement (e.g., Maccoby & Jacklin, 1974). That males have a small advantage on math and science exams whereas females tend to perform slightly better on verbal ability measures has been well documented in the literature. Less research has been directed towards the gender discrepancy in history performance, where males manifest superior scores on standardized tests of history achievement (Breland, Danos, Kahn, Kubota, & Bonner, 1994; Bridgeman & Lewis, 1994; Mazzeo, Schmitt, & Bleistein, 1993; National Assessment of Educational Progress, 1988; Willingham & Cole, 1997; Zwick & Ercikan, 1989). This disparity appears not to be due to self-selection by male versus female test-takers because the gap is also prevalent on measures in which the sample is nationally representative. Furthermore, the difference becomes more pronounced with increasing grade level. Despite this consistent male advantage, few studies have attempted to systematically explore the relationships among test scores, gender, and other variables, thereby rendering it difficult to identify the sources of gender differences in history performance.

Often, when average group differences arise, members of the lay public assume that biased test items are the culprits for the discrepancies. The existence of adverse impact, or inequality in group means, in and of itself is not a compelling reason to suspect bias. Nevertheless, it is not uncommon to find arguments delineating ways in which standardized test items are purportedly partial toward males: Questions reference males more frequently than females (Tittle, McCarthy, & Steckler, 1974), females are presented in stereotypical roles (Diamond & Tittle, 1985), item content reflects males' experiences (Walter & Young, 1997), and so forth. These issues represent serious validity and fairness concerns for developers of large-scale assessments, especially if test performance is indeed impacted by item characteristics that are unrelated to the intended construct.

This investigation examined features of the National Education Longitudinal Study of 1988 (NELS:88) multiple-choice history items that may contribute to gender differences. Several differential item functioning (DIF) analyses, varying the matching criteria, were carried out with a supplemental survey study that helped validate interpretations of underlying DIF causes. Implications were drawn for users of large-scale achievement test data and for curriculum developers.

Background

Whenever males demonstrate higher average test scores than females, questions arise concerning whether those disparities are due to actual achievement differences, bias in the test, or some combination of both. Although discrepancies in the proportions of items answered correctly are often used as evidence of bias, this argument does not allow for the possibility that performance differences are the result of real differences in the construct being assessed. To investigate bias at the item level, developers of large-scale assessments conduct a differential item functioning (DIF) analysis, which examines the relative performance of males and females while minimizing its confound with differences in ability. DIF procedures compare the probabilities of a correct response from males and females who are matched on some measure of achievement, typically total test score (Linn, 1993). Statistically significant differences in probabilities correct indicate that the particular item should be flagged for further inspection.

Once items have been identified as displaying DIF, a substantive analysis of those questions is conducted to ascertain whether they might represent artifactual occurrences of DIF. This is achieved in several ways, including investigating patterns of DIF across items with similar content or examining the magnitude of DIF in relation to item characteristics (Scheuneman & Gerritz, 1990). Such procedures are intended to identify the features that produce performance differences. Researchers must then judge whether those item attributes are relevant to the construct being assessed. If such characteristics represent construct-irrelevant item difficulty, there may be evidence of item bias within the test.

Total test score is the most commonly chosen matching variable because many standardized tests are designed to measure a single trait or unidimensional construct. For tests that in fact assess multiple constructs, however, total test score may not be the most appropriate criterion (Ackerman, 1992; Clauser, Nungester, & Swaminathan, 1996; Hamilton, 1997). When multidimensionality arises because items assessing single but different factors have been combined within a test, using subtest scores as conditioning variables may be a better alternative. Studies have shown that the number of items previously identified as displaying DIF can decrease by up to one third when examinees are matched on subtest scores as opposed to total test scores (Clauser, Mazor, & Hambleton, 1991).

In the present study, total test score and subtest scores served as the matching criteria. The rationale for matching on subtest scores stemmed from a hypothesis that the observed gender gap on history achievement tests arises primarily on items pertaining to specific content areas that reflect traditional sex-role stereotypes (Gossweiler & Slevin, 1995; Tyack & Hansot, 1990; Walter & Young, 1997). By matching on a more refined criterion, such as subscores derived from these content areas, it was believed that fewer DIF items will emerge, as item performance is compared for groups of examinees whose proficiency levels are presumably more homogenous than those matched on total test score alone.

Methodology

NELS:88 Database

The NELS:88 is a stratified, nationally representative longitudinal study that followed a sample of 24,599 8th graders into the 10th and 12th grades. The purpose of the NELS:88 tests was to assess individual status and growth in four achievement areas: math, science, reading, and history/geography/ citizenship. Attempts were made to include items that tapped general knowledge found in most curricula rather than items requiring specialized content knowledge. At each grade level, all examinees took the same history form, which consisted of 30 dichotomously scored multiple-choice items administered within a 14-minute testing period. For equating purposes, there were common history items on each form. However, to reduce ceiling effects, it was necessary to eliminate the easiest 8th-grade items in later forms and replace them with alternative items of greater difficulty. Thus, the 10th- and 12th-grade tests were grade-level adaptive, as each successive form was more difficult than the previously administered test. Due to security restrictions, actual items cannot be presented, and only brief descriptions will be given.

Creating the Subtests

Previous work with the NELS:88 math and science multiple-choice exams identified several psychologically meaningful subscores (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). To ascertain whether interpretable achievement dimensions could be derived from the multiple-choice history test, the NELS:88 history items at the 10th-grade were subjected to a full-information factor analysis. Although the data did not strictly fit the unidimensional requirements, substantive interpretability suggested that only one factor should be retained.

Even when tests appear to approximate unidimensionality, they can contain items that measure more than one skill (Clauser et al., 1991). In such cases, the utility of further grouping the items into subtests should be examined. In the present study, the subtests were constructed with the intention of exploring the hypothesis that gender-differentiated performance is related to specific content areas. Previous research has suggested that internalization of sex-role stereotypes may lead males and females to perform better on items corresponding to sex-role appropriateness (Gossweiler & Slevin, 1995; Tyack & Hansot, 1990). In other words, males perform better on traditionally “masculine” items whereas females are advantaged on stereotypically “feminine” questions.

In order to create the subtests, it was necessary to define masculine and feminine content. Masculine content typically includes themes of power, conflict, or control (Gossweiler & Slevin, 1995; Walter & Young, 1997). Prior studies have identified five topics in history as reflecting traditional masculine ideals and as areas in which boys excel: war, politics, historical documents, economics, and occupation of territories (Gossweiler & Slevin, 1995; Tyack & Hansot, 1990; Walter & Young, 1997). Although the literature does not explicitly identify any feminine historical topics, there exist areas in which females tend to show more interest. Conceivably, girls may be expected to hold a relative advantage on items pertaining to these historical content areas. These themes consist of individual liberty, equality, social consequences, religion, and food (Allport, Vernon, & Lindzey, 1970; Hansen & Campbell, 1985; Pratto, Stallworth, & Sidanius, 1997). Using these themes to guide the construction of the subtest, items pertaining to civic duties, minorities, Constitutional rights, religion and food were defined as feminine content.

Two graduate students, one male and one female, with undergraduate degrees in history, independently classified the items as masculine, feminine, both, or neither (e.g., gender-neutral). Initial interrater agreement on the 30 items was .90. Cohen's Kappa, which adjusts for the amount of agreement that would have been expected by chance alone, was .81, suggesting that interrater reliability was adequate. To resolve discrepancies, the investigator independently coded each of the three items for which the classifications assigned by the two raters differed. The classification by the investigator agreed with one of the categorizations of the two raters in every instance, so the code assigned by the investigator was used in subsequent analyses. Of the 30 items, 17 items were classified as masculine, 12 items as feminine and 1 item as neutral. There were too few items considered neutral so only masculine and feminine subtests were constructed.

It should be emphasized that there were numerous ways in which masculine and feminine content could be defined, and the criteria chosen by the present study should not be interpreted as the definitive standard. Furthermore, even with specific guidelines governing the creation of the subtests, the classification of items into the categories (e.g., masculine vs. feminine) was not always clear-cut. In general, however, previous studies have shown that carefully constructed subtest scores can provide additional achievement information that may be obscured by total test scores (Donlon, Hicks, & Wallmark, 1980).

DIF Procedures

The study applied the Mantel-Haenszel (MH) method twice to each item, once conditioning on total test score and once using the subtest to which the item had been classified (hereafter referred to as the gender subtest). A concern arose that the comparison of results from the total test score with those from the gender subtests would be confounded by test length. Without additional analyses, differences in DIF results may be attributable to differences in reliability as opposed to differences in content representation. Consequently, two control subtests were created. This was achieved by randomly assigning each of the items to one of two groups with test lengths and reliabilities approximately equal to those of the masculine and feminine subtests. A check of the content of the control subtests matched that of the total test reasonably well. The MH procedure was then applied again, with the control subtests serving as the matching variable. That is, the criterion for each item was the control subtest to which the item had been randomly assigned. This procedure is similar to that used by Clauser et al. (1991) and facilitates interpretations that compare results between total test and subtest matching.

Survey Study Procedures

To supplement the statistical analysis, a questionnaire designed to validate the item classifications was administered. The survey study was conducted with 432 high school juniors and seniors enrolled in two different schools. Students were asked to complete a subset of 20 multiple-choice items and to respond to a posttest survey that elicited their perceived sex-typing of the administered items. Specifically, they were presented with each item stem and asked to decide whether the item favored either gender. Although the students cannot be considered representative of the U.S. population of high school students, their responses help validate the definitions of masculine and feminine content.

Results

Distributions of Achievement

The descriptive statistics and representation patterns for each of the total and subtest scores are presented in Table 1. The extent to which gender differences are manifested depends upon the score measure being considered. Using total test score as the criterion, males demonstrated a small advantage. This gap became even more pronounced on the masculine subtest, where the male average was approximately one quarter standard deviation greater than the female average. On the feminine subtest, however, females demonstrated a slight advantage.

As with the descriptive statistics, the representation imbalances are consistent with those predicted by content stereotypes. There was a higher proportion of males than females classified as low achievers when the feminine subtest was the

Table 1

Descriptive Statistics and Representation Patterns on Each Score Measure (*N* = 16429)

	Males		Females		Ratio of females to males below 10th percentile	Ratio of females to males above 90th percentile
Score measure	Mean	*SD*	Mean	*SD*		
Total test	19.26 (.007)	6.13	18.37 (.007)	5.62	.98	.64
Masculine subtest	9.49 (.004)	3.68	8.51 (.004)	3.43	1.09	.59
Feminine subtest	9.03 (.003)	2.74	9.15 (.003)	2.53	.79	.88

Note. Ratio above 1.0 indicates more females than males. Standard errors given in parentheses.

criterion, but the opposite pattern appeared when using the masculine subtest as the standard. Among high achievers, the gender imbalance favored males, regardless of the outcome measure. The discrepancy was exacerbated with the masculine subtest, and was mitigated with the feminine subtest. Evidently, performance differences are most apparent on the masculine content.

DIF Results

Results were converted to a three-level classification system that reflected the degree of DIF in test items. This was achieved by rescaling the odds-ratio value. Specifically, "D" is defined as the log of the combined odds ratio multiplied by –2.35. If the absolute value of D is not significantly different from zero or is less than 1.0, the items were labeled as "A." "C" items have absolute values of D that are greater than 1.5 and are significantly greater than 1.0. Items classified as B are those that do not meet either criterion (Camilli & Shepard, 1994). "A" items are considered to be free of DIF, "B" items are not ideal questions, but are acceptable for use, and "C" items are to be replaced unless necessary for test specifications (Zwick & Ercikan, 1989). Only "B" and "C" items are items of substantive concern.

Correlations and reliabilities among the subtests and total test scores are given in Table 2. Table 3 shows the results of the DIF analyses for each item, with Type I error rate of .01. Items are organized in the table by subtest classification. The first column gives the item numbers, the second provides brief descriptions, and the third presents the subtest to which the item had been classified. *P*-values for males and females are shown in the fourth and fifth columns, respectively. The remaining three columns indicate which items showed statistically significant DIF as well as

Table 2

Correlations and Reliabilities for the Subtests and Total Test Scores at the 10th Grade

	Total test score	Masculine subtest	Feminine subtest	Control subtest 1	Control subtest 2
Total test score	(.85)				
Masculine subtest	.94	(.78)			
Feminine subtest	.88	.67	(.71)		
Control subtest 1	.95	.94	.78	(.77)	
Control subtest 2	.93	.83	.87	.77	(.73)

Note. Reliabilities are given along the diagonal.

the degree of DIF. The direction of statistically significant DIF is indicated by “M” and “F,” and the severity of DIF is given by “A”, “B,” or “C.” Items with neither “M” nor “F” had DIF that failed to reach the .01 significance level.

As shown in Table 3, the choice of criterion can influence the identification of DIF items. The total test score analysis identified 17 DIF items whereas the control subtests flagged 21 questions. In contrast, conditioning on the gender subtests flagged only 13 items. This 23% reduction in the number of items identified represents approximately 13% of the total item pool and is beyond what would be expected as a typical false positive error rate using .01 significance level.

Perhaps more important than the number of items showing statistically significant DIF is the number of items of substantive concern. Although few “B” or “C” items were identified, it was apparent that classification of an item as “B” or “C” depended upon the matching criterion. This is best exemplified by the item pertaining to grain in the diet. It had been categorized as exhibiting severe DIF (e.g., “C”) by the control subtests, as moderate DIF (e.g., “B”) by the total test score, but as essentially DIF-free (e.g., “A”) by the gender subtests. In a similar vein, both the total test score and control subtests analyses identified an item concerning the atomic bomb as a question that should be replaced; however, this same item was considered acceptable for use by the gender subtest analysis. Evidently, conditioning on the gender subtests not only reduces the number of DIF items identified, but also lessens the severity of DIF associated with particular items. Overall, the control subtests identified five items of substantive concern, the total test score flagged two questions, and the gender subtests identified only one (see Table 3).

Table 3

NELS:88 10th-Grade History Items and Descriptions

					Matching variable		
Master item #	Description of item	Subtest	Male *P*-value	Female *P*-value	Total test	Gender subtests	Control subtests
1	Manufacturing technique	M	.84	.82	A	A	A
2	Development of settlement areas	M	.68	.61	MA	A	MA
3	Great Depression	M	.64	.62	A	A	MA
4	Federal government branches	M	.59	.56	A	FA	A
6	Westward movement	M	.56	.46	MA	MA	MA
8	Watergate resignation	M	.72	.64	MA	MA	MB
10	House of Representatives	M	.55	.50	A	A	A
11	United States Senators	M	.48	.43	A	A	A
13	Unconstitutional Congressional act	M	.54	.51	A	FA	A
14	Declaration of Independence author	M	.82	.81	A	FA	A
15	Feature after 1950	M	.74	.71	MA	MA	MA
18	Treaties submitted to Congress	M	.32	.27	A	A	A
29	First atomic bomb	M	.81	.68	MC	MB	MC
32	Union membership	M	.34	.31	A	FA	A
34	Separation of colonies	M	.60	.49	MA	MA	MA
37	Stamp Act	M	.56	.55	FA	FA	FA
38	Action during Korean War	M	.51	.40	MA	MA	MA
5	Grain in diet	F	.65	.73	FB	FA	FC
7	Underground Railroad	F	.82	.84	FA	A	FA
9	Immigration during 1970s-80s	F	.56	.51	A	A	MA
16	Not a constitutional right	F	.90	.92	FA	A	FB
17	Time line showing food changes	F	.84	.85	FA	A	FA
19	Group opposing equality	F	.90	.91	A	A	A
21	Effect of voting requirements	F	.45	.44	FA	A	FA
27	Rights in Constitution	F	.78	.82	FA	FA	FB
30	Freedom of speech and religion	F	.81	.82	A	A	FA
31	Principle of lawyer appointment	F	.66	.67	FA	A	FA
33	Brown v. Board of Education	F	.59	.61	FA	A	FA
35	Goal of United Nations	F	.69	.70	FA	A	FA
39	Social Security system	N	.40	.40	A	A	FA

Note. M indicates DIF in favor of males, $p < .01$. F indicates DIF in favor of females, $p < .01$. Columns with neither M nor F indicate DIF failed to reach statistical significance. The severity of DIF is given by “A,” “B,” or “C.”

There was a high degree of correspondence between the items identified by total test score and those flagged by the control subtests, such that all of the total test score DIF items were a subset of the control subtests DIF items. When examinees were matched on the gender subtests, there was a substantial shift in the specific items identified. In other words, items that had been flagged as DIF when total test score was the criterion were seldom identified as such when matching on the gender subtests. This trend was observed primarily with the feminine items.

To better understand the nature of the DIF items, the questions favoring males and females were analyzed. As shown in Table 4, there was a strong tendency for performance to be consistent with sex-role appropriateness when total score was the criterion. Approximately 41% of the masculine items were conditionally easier for males, whereas 75% of the feminine items were conditionally easier for females. In contrast, the DIF items identified by the gender subtests were less differentiated along content stereotypic lines, with 35% of masculine items favoring males and 17% of feminine items favoring females (see Table 4). Somewhat unexpectedly, conditional on the masculine subtest score, females demonstrated superior performance on many masculine items, particularly those pertaining to Congress. However, performance on the single item flagged as potentially biased (atomic bomb) remained in the gender stereotypic direction.

As a limited safeguard against the inherent circularity of the DIF indices, several other analyses were conducted. It has been recommended that DIF statistics be computed iteratively to obtain a “purified” matching criterion. That is, if items identified as displaying DIF are removed in the first stage and the analysis is repeated, results in the latter stages are less contaminated by the effects of potentially

Table 4

DIF Items Favoring Each Gender by Item Type and Matching Variable

	Matching criterion			
	Total test score		Gender subtest score	
Item type	Number of DIF items favoring males	Number of DIF items favoring females	Number of DIF items favoring males	Number of DIF items favoring females
Masculine items (*N* = 7)	7	1	6	5
Feminine items (*N* = 12)	0	9	0	2
Neutral items (*N* = 1)	0	0	0	0

biased items. It is important, however, that the anchor set of items not be achieved at the expense of distortions in the test content domain. Camilli and Shepard (1994) have suggested that judgment be used in conjunction with statistical indices, so that an item is removed only if it can be interpreted as showing bias. In the present study, only “B” and “C” items were of substantive concern. Purified anchor sets were then created by eliminating the identified “B” and “C” items within a given matching criterion. The analyses were repeated with the purified anchor sets serving as the matching standards. Results did not change dramatically from those reported earlier.

An analysis of the chosen distractors was also conducted. On the item concerning the atomic bomb, females were approximately 13% more likely than males to choose the incorrect option that included Germany rather than the correct answer that included Japan. Females were also 8% more likely to incorrectly identify the purpose of the Declaration of Independence, instead choosing the “Articles of Confederation” as the correct answer. On an item asking about grain in the world’s diet, 19% of males, compared to only 7% of females, gave the incorrect response, “grain is easier to prepare.” Overall, however, there was little evidence that males and females were being lured by particular multiple-choice options.

Survey Study

An examination of the DIF items can reveal characteristics that contribute to discrepancies between boys’ and girls’ test scores. In the present research, substantive judgment suggested that performance on the DIF items was related to the subtests corresponding to sex-role stereotypes. Inspection of the items alone, however, cannot provide an external validation for the manner in which the subtests were created. The survey study was intended to address this concern by evaluating the plausibility of the definitions that guided the subtests construction.

Of the 20 multiple-choice items included in the survey study, 12 had been categorized as masculine and 8 as feminine. To validate the categorizations, 432 high school juniors and seniors were asked whether they believed a particular item would show performance differences in favor of one of the sexes. On any given item, between 21% and 54% of the responses indicated sex typing existed. Among those who perceived an item to be gender-typed, responses corroborated the subtest classifications in all but one instance. That is, if an item had been classified as masculine, at least 75% of the sex-typed responses expected males to manifest better performance. An analogous result was found with the feminine classifications. The one item in which students' responses differed from the prior categorization concerned immigration during the 1970s and 1980s. It had been classified as feminine, but the majority of the sex-typed responses perceived a male advantage. Notably, in the national sample, this was the single feminine item that had shown DIF in favor of boys. The anomalous item notwithstanding, the results provided some empirical support for the manner in which the subtests were defined and created.

Discussion and Implications

In comparison to total test score, gender subtest matching flagged fewer DIF items and lessened the severity of DIF. If this finding were solely attributable to shorter test length or less reliability, conditioning on the control subtests should have yielded similar results. However, as control subtest matching did not lead to a commensurate decrease in either the severity of DIF or number of DIF items identified, it appears that the differences between the total test score and gender subtest results are due to differences in content representation. Specifically, the conditional difference in difficulty on the NELS:88 history test may stem from differentiated historical knowledge corresponding to sex-role appropriateness.

Why might conditioning on gender subtest scores lead to a reduction in DIF questions? One possible explanation is that most of the studied items measured one or more abilities that were sufficiently captured by the subtest to which they had been classified. If the probability of answering an item correctly depends only on the given subtest, then responses to other items not included in that particular subtest do not add any useful classification information. In fact, such responses are likely to add error, which lead to less homogenous strata, and hence, an increase in the number of DIF items identified. This may explain why the control subtests and total test scores flagged more DIF items than did the gender subtest scores, as the two former criteria included responses that were irrelevant to the latent ability space accounting for the item of interest. Presumably, item performance was compared for groups of examinees whose proficiency levels were more similar under the gender subtest analysis than under either the total test scores or control subtests analyses.

The results underscore the importance of considering other measures of achievement besides total test score. An analysis in which total score is the only criterion would have obscured details of the particular areas in which males excel.

Both the descriptive statistics and the DIF results suggest that the male advantage on this particular test is limited to specific content areas, and is not a result of overall superiority in history. It appears that under some circumstances, considering subsets of items may elucidate important details about the relationships of achievement and performance.

For test developers, one of the first important decisions in designing a test is the choice of content to be included. As shown in this investigation, the kinds of items selected can have substantial influence not only on the extent to which average gender differences are demonstrated, but also on the gender imbalances with respect to admission, placement, and scholarship decisions. In the present study, it is possible to narrow the observed discrepancies by limiting the proportion of masculine items included on the test. However, knowledge about wars, politics, and other masculine content represents a legitimate aspect of history ability and should not be excluded from the sample of skills without an adequate reason. That is, the logic for replacing items that may contribute to the observed differences should not be derived from an attempt to manipulate the magnitude of the gender gap. Instead, questions should be eliminated only if they do not meet the test specifications or are irrelevant to the intended test construct.

If the test items are considered valid questions that denote real differences in attainment between boys and girls, then avenues other than changes in test content need to be explored in order to achieve equitable outcomes. One likely possibility is that the performance discrepancies reflect differences in opportunities to learn. Indeed, an analysis of the NELS:88 data indicated that boys were more likely than girls to have had exposure to history courses. Perhaps females were not choosing to enroll in history at the same rate as males because of the dearth of historical female figures within the curriculum, which is thought to adversely affect girls' level of engagement (Belenky, McVicker, Goldberger, & Mattuck, 1986; Lerner, 1975; Noddings, 1992). Studies have suggested that females may be less inclined than males to show interest in history because it does not realistically reflect their experiences, and does not allow for perspectives other than those of males (Cherryholmes, 1983; Ferree & Wienand, 1987; Hannam, 1993). It appears that revising history instruction to encompass females' interests and points of views may increase their involvement within history, and may help narrow the gender gap on measures of history achievement.

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.

Allport, G. W., Vernon, P. E., & Lindzey, G. (1970). *Study of Values—A scale for measuring the dominant interests in personality* (3rd ed.). New York: Houghton Mifflin.

Belenky, M. F., McVicker B. C., Goldberger, N. R., & Mattuck, J. T. (1986). *Women's ways of knowing*: *The development of self, voice, and mind*. New York: Basic Books.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement History examination. *Journal of Educational Measurement*, *31,* 275-293.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*, 37-50.

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

Cherryholmes, C. (1983). Knowledge, power, and discourse in social studies education. *Journal of Education*, *165,* 341-358.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement*, *15*, 353-359.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33,* 453-464.

Diamond E., & Tittle, C. (1985). Sex equity in testing. In S. Klein (Ed.), *Handbook for achieving sex equity* (p. 167). Baltimore, MD: John Hopkins University Press.

Donlon, T. F., Hicks, M. M., & Wallmark, M. M. (1980). Sex differences in item responses on the Graduate Record Exam. *Applied Psychological Measurement*, *4*, 9-20.

Ferree, M. M., & Wienand, A. (1987). *Does ignorance breed contempt? The balanced curriculum. Fall 1986.* Unpublished report, Women's Studies Program, University of Connecticut.

Gossweiler, R. S., & Slevin, K. F. (1995). The importance of gender in the assessment of historical knowledge. *Research in Higher Education, 36*, 155-175.

Hamilton, L. S. (1997). *Identifying differential item functioning on science achievement tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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.

Hannam, J. (1993). Women, history, and protest. In D. Richardson & V. Robinson (Eds.), *Thinking feminist* (pp. 302-323), New York: The Guilford Press.

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII. Strong-Campbell Interest Inventory—Form T325 of the Strong Vocational Interest Blank* (4th ed.). Stanford, CA: Stanford University.

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, 32*, 124-150.

Lerner, G. (1975). Placing women in history: Definitions and challenges. *Feminist Studies, 3*, 5-14.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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; ETS Research Report No. 93-5). New York: College Entrance Examination Board.

National Assessment of Educational Progress. (1988). *The History Report Card: Trends and achievement based on the 1988 National Assessment*. Princeton, NJ: Educational Testing Service.

Noddings, N. (1992). Social studies and feminism. *Theory and Research in Social Education, 20,* 230-241.

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.

Pratto, F., Stallworth, L. M., & Sidanius, J. (1997). The gender gap: Difference in political attitudes and social dominance orientation. *British Journal of Social Psychology, 36,* 49-68.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

Tittle, C. K., McCarthy, K., & Steckler, J. F. (1974). *Women and educational testing*. Princeton, NJ: Educational Testing Service.

Tyack, D. B., & Hansot, E. (1990). *Learning together*. New Haven: Yale University Press.

Walter, C., & Young, B. (1997). Gender bias in Alberta Social Studies 30 Examinations: Cause and effect. *Canadian Social Studies, 31*(2), 83-86, 89.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

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