Journal of Educational Measurement
Spring 1989, Vol. 26, No. 1, pp. 55–66

Analysis of Differential Item Functioning in the NAEP History Assessment

Rebecca Zwick
Educational Testing Service
and
Kadriye Ercikan
Stanford University

The Mantel-Haenszel approach for investigating differential item functioning was applied to U.S. history items that were administered as part of the National Assessment of Educational Progress. On some items, blacks, Hispanics, and females performed more poorly than other students, conditional on number-right score. It was hypothesized that this resulted, in part, from the fact that ethnic and gender groups differed in their exposure to the material included in the assessment. Supplementary Mantel-Haenszel analyses were undertaken in which the number of historical periods studied, as well as score, was used as a conditioning variable. Contrary to expectation, the additional conditioning did not lead to a reduction in the number of DIF items. Both methodological and substantive explanations for this unexpected result were explored.

The National Assessment of Educational Progress (NAEP) is a survey of the academic achievements of American students that began in 1969. The Mantel-Haenszel (MH), 1959, approach to differential item functioning (DIF) analysis developed by Holland and Thayer (1988) was applied to U.S. history items that were administered in 1986 as part of a project supported by NAEP and the National Endowment for the Humanities (see Applebee, Langer, & Mullis, 1987). On about 30 percent of the items, there was some evidence that either blacks, Hispanics, or females performed more poorly than other students, conditional on number-right score.

It was hypothesized that this could have resulted, in part, from the fact that ethnic and gender groups differed in their exposure to the material included in the history assessment. In this study, the results of a standard Mantel-Haenszel DIF analysis are compared to results obtained from supplementary analyses in which history course background, as well as score, is used as a conditioning variable. The purpose of this more refined matching procedure is to achieve a situation in which item performance is compared for groups of students who are of similar overall proficiency and have been exposed to similar curricula. If the original findings were indeed a reflection of differences in curriculum exposure, the new analyses should produce fewer DIF items.

The authors thank Paul Holland for consultation and Jennifer Nelson and Laurie Barnett for statistical programming assistance. A portion of this work was conducted while the second author was a predoctoral fellow at ETS.

The U.S. History Assessment

History items were included in 4 of the 92 booklets administered to a national sample of students who were 17 years old or in grade 11 in the 1986 NAEP assessment. Each of the 4 booklets contained one of four history blocks (H1, H2, H3, or H4), as well as a block of literature items and a block of reading items. The objectives for the history assessment, as well as the items themselves, were developed through consultation with a committee of U.S. history specialists. Potential items were then reviewed by more than 50 educators from around the country. Each U.S. history block consisted of 34 to 36 cognitive items and a common set of history background items that included questions about previous courses in history. The four history blocks were constructed to be parallel in content and yielded similar item analysis results, although block H1 was somewhat easier than the remaining three blocks (see Table 1). The students who took each of the four blocks were random samples from the same population. As in all NAEP assessments, no results were reported at the individual student level.

For reporting the history results, NAEP used item response theory methods to derive a scale based on the responses of the 7,812 students who were in grade 11. DIF analyses were based on the responses of 7,743 11th graders; students who failed to answer any items or who received defective test booklets were excluded.

In interpreting the results described here, it is necessary to consider that NAEP collects data using a stratified multistage cluster sampling scheme in which students have differential probabilities of selection. As in most surveys, each respondent is assigned a sampling weight. Based on preliminary investigation, it appears that the NAEP sampling weights have little impact on the Mantel-Haenszel delta difference (MH D-DIF) statistic (equation 7). Because of cluster effects, however, the distributions of MH D-DIF and Mantel-Haenszel chi-square (MH CHISQ) statistic (equation 3) will differ from their distributions under simple random sampling. In the analyses described here, no adjustment was made for the complex sampling scheme. Therefore, the significance probabilities (p-values) discussed in the following sections can be assumed to depart to some degree from the actual significance probabilities. The classification of items into A, B, and C categories could also be affected. The focus of the present study, however, is the comparison of two competing analysis methods: 1) conditioning on score only and 2) conditioning on both score and history course background.

Table 1
NAEP History Assessment:
Descriptive Statistics

Block	Numbers of Items	KR-20 Reliability	Average Tetrachoric	Mean	S.D.	Mean p
H1	36	.84	.39	20.8	6.3	.58
H2	36	.83	.35	19.2	6.4	.53
H3	35	.82	.40	16.9	6.1	.48
H4	34	.87	.48	19.2	6.9	.57

Note. For each block, the sample size was approximately 1950.

Table 2
Sample Sizes for DIF Analyses

	Male	Female	White	Black	Hispanic
H1	964	989	1375	321	198
H2	945	984	1365	330	168
H3	935	975	1346	306	201
H4	1018	933	1410	308	185

Note. Six examinees were excluded from Analysis 2 because they were missing information on historical periods studied. Students who failed to reach an item were excluded from the DIF analysis for that item.

Analysis 1: Conditioning on Score Only

Within each of the four history blocks, DIF analyses were conducted to compare the performance of males and females, whites and blacks, and whites and Hispanics, conditional on number-right score. The sample sizes for each group are given in Table 2.

The standard Mantel-Haenszel (1959) approach to DIF analysis, developed by Holland and Thayer (1988), involves the creation of K two-by-two tables, where K is the number of score categories. Because there were few examinees at the lower end of the distribution, scores 0–6 and scores 7–9 were collapsed. This collapsing scheme was selected over other possible schemes because it minimized the number of unmatched focal group members. For the k^{th} score level, the data can be displayed as in Table 3. Here, F denotes the focal group (blacks, Hispanics, and females, respectively, in the analyses considered here) and R denotes the reference group. The numbers of examinees in the R and F groups are denoted by n_{Rk} and n_{Fk}, respectively; m_{1k} represents the number of examinees who answered the item correctly and m_{0k} is the number who answered incorrectly. A_k and C_k denote the numbers of examinees in the R and F groups, respectively, who answered correctly; B_k and D_k are the numbers of examinees in the R and F groups who answered incorrectly. T_k is the total number of examinees. (For both Analysis 1 and Analysis 2, examinees who did not reach an item were excluded from the DIF analysis for that item. This eliminates problems in interpretation that can result when the focal groups and reference groups have different rates of completing the item block.)

Table 3
Data for the k^{th} Matched Set of Reference and Focal Group Members

	Score on Studied Item		
Group	1	0	Total
R	A_k	B_k	n_{Rk}
F	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

As described in Holland and Thayer (1988), it is assumed that, within each stratum, data for the R and F groups have been acquired by obtaining (simple) random samples of fixed sizes (n_{Rk} and n_{Fk}) from pools of reference and focal group members. A_k and C_k are then independent binomial random variables with parameters (n_{Rk}, p_{Rk}) and (n_{Fk}, p_{Fk}), respectively. In the present context, p_{Rk} represents the probability of answering the item correctly for members of the reference group in the k^{th} stratum; p_{Fk} is the corresponding probability for the focal group. We wish to test the hypothesis

$$H_0: \quad \frac{p_{Rk}/q_{Rk}}{p_{Fk}/q_{Fk}} = 1, \quad k = 1, 2, \ldots K, \qquad [1]$$

versus

$$H_1: \quad \frac{p_{Rk}/q_{Rk}}{p_{Fk}/q_{Fk}} = \omega, \quad \omega \neq 1. \qquad [2]$$

The parameter ω represents the common odds ratio for the K 2×2 tables. The uniformly most powerful unbiased test of H_0 versus H_1 is provided by the Mantel-Haenszel chi-square statistic:

$$\text{MH CHISQ} = \frac{\left(\left| \sum_k A_k - \sum_k E(A_k) \right| - \frac{1}{2} \right)^2}{\sum_k \text{Var}\,(A_k)} \qquad [3]$$

where

$$E(A_k) = n_{Rk} m_{1k} / T_k \qquad [4]$$

and

$$\text{Var}\,(A_k) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}. \qquad [5]$$

The statistic in [3] has a chi-square distribution with one degree of freedom when the stated assumptions are met and H_0 is true. Mantel and Haenszel provided the following estimator of ω:

$$\hat{\omega}_{MH} = \frac{\Sigma A_k D_k / T_k}{\Sigma B_k C_k / T_k}. \qquad [6]$$

At Educational Testing Service (ETS), the statistic typically used as an index of differential item performance is

$$\text{MH D-DIF} = -2.35 \ln (\hat{\omega}) \qquad [7]$$

(see Holland & Thayer, 1988). Using the preceding formulation will result in negative values of MH D-DIF for items that favor the reference group and positive values for items that favor the focal group.

The following rules have been developed for use by ETS testing programs in interpreting the results of DIF analyses:

"A" items are those for which MH D-DIF is not significantly different from 0

Table 4
Results of Male-Female Analyses:
Numbers of A, B, and C Items

		Analysis 2						
		Male +			Female +			
Analysis 1		A	B	C	A	B	C	Total
Male +	A	46	2	0	3	0	0	51
	B	0	12	0	0	0	0	12
	C	0	1	3	0	0	0	4
Female +	A	1	0	0	58	1	0	60
	B	0	0	0	1	12	0	13
	C	0	0	0	0	0	1	1
Total		47	15	3	62	13	1	141

Note. The labels "Male +" and "Female +" indicate which group showed superior conditional performance on the corresponding items.

(α = .05) or has an absolute value less than 1. These items are considered to be free of DIF.

"B" items are those for which MH D-DIF is significantly different from 0 (α = .05) and has either (a) an absolute value at least 1 but less than 1.5 or (b) an absolute value at least 1 but not significantly greater than 1 (α = .05). These items may be used, but if there is a choice among otherwise equivalent items, it is considered desirable to select for inclusion in a test those with the smallest absolute value of MH D-DIF.

"C" items are those for which the absolute value of MH D-DIF is at least 1.5 and is significantly greater than 1 (α = .05). These items are to be selected only if it is essential to meet test specifications.

For purposes of this study, the NAEP U.S. history items were classified into A, B, and C categories. Results were tabulated separately for items that favored the reference group (conditional on score) and those that favored the focal group. The right margins of Tables 4, 5, and 6 show the numbers of DIF items for Analysis 1 according to this classification system.

For example, the right margin of Table 4 shows that in Analysis 1, the

Table 5
Results of White-Black Analyses:
Numbers of A, B, and C Items

		Analysis 2						
		White +			Black +			
Analysis 1		A	B	C	A	B	C	Total
	A	50	0	0	0	0	0	50
White +	B	0	14	1	0	0	0	15
	C	0	0	0	0	0	0	0
	A	3	0	0	61	1	0	65
Black +	B	0	0	0	2	6	0	8
	C	0	0	0	0	0	3	3
Total		54	14	1	63	7	3	141

Note. The labels "White +" and "Black +" indicate which group showed superior conditional performance on the corresponding items.

Table 6
Results of White-Hispanic Analyses: Numbers of A, B, and C Items

		Analysis 2						
		White +			Hispanic +			
Analysis 1		A	B	C	A	B	C	Total
White +	A	46	4	0	6	0	0	56
	B	0	14	1	0	0	0	15
	C	0	0	0	0	0	0	0
Hispanic +	A	1	0	0	58	0	0	59
	B	0	0	0	7	3	0	10
	C	0	0	0	0	0	1	1
Total		47	18	1	71	3	1	141

Note. The labels "White +" and "Hispanic +" indicate which group showed superior conditional performance on the corresponding items.

male-female comparison yielded 51 + 12 + 4 = 67 items for which MH D-DIF was negative, indicating that males performed better on the item, conditional on score. Of these items, 51 were A items and thus not of concern, 12 were B items, and 4 were C items. On 74 items, the conditional performance of females was better. These items included 60 A, 13 B, and 1 C item.

Tables 7, 8, and 9 show the results obtained if only the statistical significance of the chi-square values is considered in classifying items. For example, of the 67 items with negative values of MH D-DIF in Analysis 1, Table 7 shows that 25 were statistically significant at $\alpha = .01$. These tables, as well as the Analysis 2 results, are discussed in later sections.

Although the results of Analysis 1 were not always interpretable with respect to item content and type, certain meaningful patterns were evident, particularly with regard to the C items.

First, consider the male-female analyses. All four C items that were easier for

Table 7
Results of Male-Female Analyses: Numbers of Items With Chi-Square Not Significant/Significant at $\alpha = 0.01$

		Analysis 2				
		Male +		Female +		
Analysis 1		Not sig.	Sig.	Not sig.	Sig.	Total
Male +	Not Sig.	32	7	3	0	42
	Sig.	0	25	0	0	25
Female +	Not Sig.	1	0	52	0	53
	Sig.	0	0	1	20	21
	Total	33	32	56	20	141

Note. The labels "Male +" and "Female +" indicate which group showed superior conditional performance on the corresponding items.

Table 8
Results of White-Black Analyses:
Numbers of Items With Chi-Square
Not Significant/Significant at $\alpha = 0.01$

		Analysis 2				
		White +		Black +		
Analysis 1		Not sig.	Sig.	Not sig.	Sig.	Total
	Not Sig.	33	13	0	0	46
White +	Sig.	0	19	0	0	19
	Not Sig.	3	0	59	0	62
Black +	Sig.	0	0	5	9	14
	Total	36	32	64	9	141

Note. The labels "White +" and "Black +" indicate which group showed superior conditional performance on the corresponding items.

males, conditional on score, pertained to World War I or World War II; two of these asked for dates. Among the 12 B items that were conditionally easier for males, 3 were also about war and 7 asked for dates.

Only a single C item that was conditionally easier for females was found: an item asking who was the inventor of the telephone! Of the 13 B items that were conditionally easier for females, 4 pertained to slavery or segregation and 2 were about women's voting rights.

In the white-black analysis, there were no C items that were conditionally easier for whites. The 15 B items that were conditionally easier for whites included 7 items involving map reading and 4 items on World War II. The three C items on which blacks performed better than whites, conditional on score, were about Martin Luther King, Harriet Tubman, and the Underground Railroad. The eight B items that were conditionally easier for blacks included two on slavery, three on the civil rights movement, and one on women's rights.

Table 9
Results of White-Hispanic Analyses:
Numbers of Items With Chi-Square
Not Significant/Significant at $\alpha = 0.01$

		Analysis 2				
		White +		Hispanic +		
Analysis 1		Not sig.	Sig.	Not sig.	Sig.	Total
	Not Sig.	36	21	6	0	63
White +	Sig.	0	8	0	0	8
	Not Sig.	1	0	62	0	63
Hispanic +	Sig.	0	0	4	3	7
	Total	37	29	72	3	141

Note. The labels "White +" and "Hispanic +" indicate which group showed superior conditional performance on the corresponding items.

Table 10
White-Black Analysis of Map Items

	DIF	Not DIF	Total
Map	7	5	12
Not Map	8	121	129
Total	15	126	141

Note. B items on which whites performed better than blacks, conditional on score

In the white-Hispanic analysis, there were again no C items that were conditionally easier for whites. The single C item on which the performance of Hispanics exceeded that of whites, conditional on score, was an item about Latin American and Asian immigration to the United States in the 1970's and 1980's. The 10 B items that were conditionally easier for Hispanics included another item about immigration, an item requiring identification of the part of the U.S. that fought for independence from Mexico, an item about Lincoln, and an item about the Emancipation Proclamation. Oddly enough, however, the 15 B items that were conditionally easier for whites than for Hispanics included 3 items on slavery or segregation and 1 item on the increase in women in the work force during World War II.

Two of the findings mentioned above involve item type rather than item content: the superior performance of whites over blacks on map items and males over females on date items. To explore these results further, consider the results displayed in Tables 10 and 11. This type of analysis allows us to see that about 58% of map items were conditionally easier for whites than for blacks, compared to only about 6% of nonmap items. About 30% of date items were conditionally easier for males, compared to about 6% of nondate items. In some cases, classification of items into categories (e.g., war vs. nonwar items) is not clearcut. In general, however, constructing tables of this kind is helpful in determining the relevance of item type or content to DIF status.

Analyses of the distractors chosen by each demographic group were conducted to explore the reasons for DIF in greater detail. In general, however, there was no evidence that the group with lower conditional performance was being lured by any particular distractor. An exception is the item on Martin Luther King. When asked what event marked King's achievement of national prominence, 25% of whites, compared to only 8% of blacks, gave the incorrect response, "Brown vs. Board of Education case in 1954."

Table 11
Male-Female Analysis of Date Items

	DIF	Not DIF	Total
Date	9	21	30
Not Date	7	104	111
Total	16	125	141

Note. B and C items on which males performed better than females, conditional on score

Analysis 2: Conditioning on Both Score and Number of Historical Periods Studied

As part of the history assessment, students were asked to indicate whether they had studied, since grade 9, the following periods of American history, which were included in the assessment: Exploration, Revolutionary War–War of 1812, Territorial Expansion–Civil War, Reconstruction–World War I, World War I–World War II, and World War II–present. Students were classified according to the number of historical periods they claimed to have studied. The number of historical periods studied (hereafter called Periods Studied) had a strong relation to overall performance on the NAEP history assessment. The first two lines of Table 12 show the estimated percent of 11th graders in the nation associated with each level of Periods Studied, along with the mean history scale value for each level. Standard errors of means and percents are given in parentheses. The remainder of the table gives the corresponding information for males, females, whites, blacks, and Hispanics.

The history scale values have a mean of 285 and a standard deviation of 40 for the 11th grade sample. It is clear that Periods Studied is strongly associated with the history scale values. For the total sample, the difference in history means between those who had studied 0–2 periods and those who had studied 6 was more than three-quarters of a standard deviation. Also, the distribution for whites differed from those of blacks and Hispanics, particularly in the tails. For instance, whereas 32% of whites had studied all 6 periods, only 24% of blacks and 22% of Hispanics had done so. The distributions for males and females were quite similar, although males were somewhat more likely to have studied all 6 periods. The rationale for Analysis 2 was that, by conditioning on Periods Studied as well as score, examinees would be more closely matched. It was expected that this more refined conditioning would produce a smaller number of items showing DIF in favor of the reference groups.

Table 12
Distribution of Number of Historical Periods Studied and History Scale Means

		Number of Historical Periods Studied				
	Sample Size	0-2	3	4	5	6
Total	7764	10.1 (0.7)	14.4 (0.7)	21.0 (0.6)	24.0 (1.1)	30.5 (0.9)
		263.0 (2.3)	277.4 (1.7)	283.0 (1.7)	288.0 (1.9)	295.1 (1.6)
Male	3875	10.1 (0.9)	14.5 (0.9)	19.5 (0.7)	23.6 (1.1)	32.2 (1.2)
		268.3 (3.1)	283.6 (1.8)	288.1 (2.1)	293.0 (2.4)	301.5 (1.9)
Female	3889	10.0 (0.9)	14.3 (0.8)	22.5 (0.8)	24.4 (1.2)	28.8 (1.1)
		257.4 (2.4)	270.9 (2.2)	278.4 (2.1)	283.0 (1.8)	287.6 (1.7)
White	5507	9.0 (0.8)	13.9 (0.9)	20.2 (0.6)	24.5 (1.4)	32.3 (1.0)
		270.1 (2.9)	283.0 (2.0)	289.1 (1.8)	293.0 (2.3)	299.5 (1.7)
Black	1273	13.0 (1.0)	16.0 (1.2)	23.6 (1.8)	23.5 (1.1)	24.0 (2.0)
		248.6 (3.2)	258.3 (2.2)	259.6 (1.9)	268.4 (2.6)	272.2 (2.5)
Hispanic	755	16.1 (1.5)	17.1 (2.4)	23.8 (1.8)	20.8 (1.6)	22.1 (2.2)
		247.4 (3.7)	262.1 (5.0)	259.8 (2.9)	268.1 (2.1)	270.1 (3.9)

Note. For each category of examinees, the first line shows the estimated percent of eleventh graders in the nation corresponding to each level of periods studied. The second line shows the history means on a scale with a mean of 285 and a standard deviation of 40. Standard errors of percents and means are given in parentheses.

In conducting Analysis 2, the collapsing scheme for score was the same as in Analysis 1. Periods Studied was grouped into five categories: 0–2, 3, 4, 5, and 6. For each history block, the number of stratification levels for Analysis 2 was, therefore, five times the number of levels for Analysis 1.

The results of Analysis 2 are given in the lower margins of Tables 4–9. In Tables 4, 5, and 6, which display the A, B, and C classifications, the results of Analysis 2 were nearly identical to those of Analysis 1. That few items changed classifications can be observed by noting that most of the off-diagonal elements are zeroes. Only in the white-Hispanic analyses were there some larger shifts and these were in the opposite direction to the predicted change: The number of items that were conditionally easier for whites increased and the number of items that were conditionally easier for Hispanics decreased. Tables 7, 8, and 9 show results that are even more surprising: If only the statistical significance of the chi-square values was considered in classifying items, all three group comparisons yielded an increase in the number of items that showed DIF in favor of the reference group (see the cell in the first row and second column) and a decrease in the items that showed DIF in favor of the focal group (see the cell in the fourth row and third column). The most dramatic change was in the white-Hispanic analysis, in which 21 items that were conditionally easier for whites, but were not statistically significant in Analysis 1, became statistically significant in Analysis 2. Two basic questions were raised by these results: (a) Why did the classification of items as A, B, or C remain relatively constant, while the classification by statistical significance showed a substantial change between Analysis 1 and Analysis 2? (b) Why did the more refined matching produce at least as many items favoring the reference group as the original analysis, regardless of which classification method was used? Substantive and technical aspects of these questions are addressed in the next two sections.

Substantive Issues

What substantive phenomena might explain the unexpected finding that additional conditioning variables did not lead to a reduction in the number of DIF items? One possibility is that Periods Studied did not add any useful classification information, given score. If the probability of answering the history items correctly were independent of Periods Studied, given score, then Analysis 2 would be expected to produce the same results as Analysis 1, as was indeed the case in terms of the A, B, and C classifications. However, examination of the joint distribution showed that Periods Studied was not redundant with score. The Pearson correlations between the two variables were approximately .20 in each of the four history blocks.

Could any nontechnical explanation account for an *increase* in the number of DIF items? One possibility is that "studying a topic" meant different things for different demographic groups. For instance, if the instruction to which minority students have access is inferior, in general, to that to which whites have access, perhaps coverage of a topic is more likely to be inadequate for minorities. If this hypothesis were true, "matching" on Periods Studied could have produced strata that were less homogeneous than the strata of Analysis 1. This hypothesis would

not seem to apply to male-female comparisons, however. A related hypothesis is that the demographic groups differed in their interpretation of the question about periods studied. It is possible that students who, in fact, had the same course background, nevertheless, responded differently to the question about periods studied and that these response tendencies were related to gender or ethnicity. If this were true, it would again be the case that the "matching" of Analysis 2 would not have resulted in greater within-stratum homogeneity.

Technical Issues

It might be hypothesized that the classification of items as A, B, and C obscured a difference in results between Analyses 1 and 2. To explore this hypothesis, three DIF statistics were examined in detail: MH D-DIF, MH D-DIF divided by its standard error (SE), see Phillips & Holland, 1987, and MH CHISQ. For each of these three statistics, the 141 values from Analysis 2 were regressed on those from Analysis 1, yielding the following results:

Statistic	Slope	Intercept	Correlation
MH D-DIF	1.0	0.0	.98
MH D-DIF/SE	1.0	0.0	.99
MH CHISQ	0.8	1.2	.92

Clearly, only the MH CHISQ values differed across the two analyses. This is somewhat disconcerting because the chi-square test has a more rigorous theoretical basis than the Mantel-Haenszel odds ratio estimator in Equation 6. One possible reason for the chi-square findings is that the complex sampling scheme has a differential effect on the two analysis methods. A more likely explanation is that the sparser tables of Analysis 2 cause the chi-square approximation to deteriorate. Each of these possibilities is discussed below.

Possible Differential Effect of Complex Sampling on Analyses 1 and 2

The effect of NAEP's complex sampling scheme on the distribution of MH CHISQ will depend upon the relation between the variables used for conditioning in the Mantel-Haenszel test and the variables used for defining clusters and strata in the sampling plan. Therefore, there is some possibility that the impact of complex sampling on the distribution of MH CHISQ could differ for Analyses 1 and 2. Further study of the distribution of MH D-DIF and MH CHISQ under complex sampling is under way.

Possible Deterioration of Chi-Square Approximation in Analysis 2

At present, the most likely explanation for the discrepancy between the two analysis methods in the number of significant items is that the distribution of MH CHISQ is affected by the pattern of sparseness that occurs in the 2×2 tables of Analysis 2. The fact that the discrepancy is largest for the white-Hispanic analysis, which has the smallest sample size, seems to support this explanation. A simulation was conducted to determine whether the chi-square findings reflected meaningful information about the Periods Studied variable or

whether they were artifactual. Using the actual data from the male-female analysis, the males at each score level were randomly allocated to an arbitrary stratification variable in such a way as to duplicate the joint distribution of score and Periods Studied; this process was repeated for females. A Mantel-Haenszel analysis was then conducted, producing results that were nearly identical to those of Analysis 2. The table showing the association between the Analysis 1 and simulation results closely resembled Tables 4 and 7. Five replications of the simulation were performed, yielding essentially the same results. Further investigations of this phenomenon are in progress.

Should Conditioning Variables In Addition to Score Be Used?

It seems that in many applications, it would be desirable to judge as problematic only those items that show DIF for groups that have been equated on measures of course background, as well as ability. However, several drawbacks should be considered:

1. Additional conditioning measures may not be readily available.
2. Adding conditioning variables may not increase within-stratum homogeneity, either because the available measures are so highly correlated with score that they do not contribute additional information, or because they are subject to errors of the kind described in this paper.
3. The sparser tables that result from multivariate matching may affect the properties of the Mantel-Haenszel chi-square.

In any case, the results of this study indicate that conditioning on additional variables within the Mantel-Haenszel framework does not necessarily decrease the number of items identified as having DIF.

References

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). *Literature & U.S. history: The instructional experience and factual knowledge of high school juniors* (NAEP Rep. No. 17–HL–01). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

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.

Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43,* 425–431.

Authors

REBECCA ZWICK, Research Scientist, Division of Statistical and Psychometric Research, Educational Testing Service, Princeton, NJ 08541. *Degrees:* BA, Antioch College; MA, PhD, University of California, Berkeley. *Specializations:* psychometric methods, applied statistics.

KADRIYE ERCIKAN, Graduate Student, Stanford University, Stanford, CA 94305. *Degrees:* BA, Essex University; MA, Stanford University. *Specializations:* educational measurement, applied statistics, operations research.