The Alberta Journal of Educational Research *Vol. 56, No. 4, Winter 2010, 459-469*

Barnabas C. Emenogu
Olesya Falenchuk

and

Ruth A. Childs
Ontario Institute for Studies in Education, University of Toronto

The Effect of Missing Data Treatment on Mantel-Haenszel DIF Detection

Most implementations of the Mantel-Haenszel differential item functioning procedure delete records with missing responses or replace missing responses with scores of 0. These treatments of missing data make strong assumptions about the causes of the missing data. Such assumptions may be particularly problematic when groups differ in their patterns of nonresponse. Using both real data from Ontario students taking the English- and French-language versions of two large-scale assessments and simulated data, this study compared the Mantel-Haenszel DIF indices produced using a total score or a proportion score as matching variables and treating missing data by listwise deletion, analysiswise deletion, or scoring missing data as incorrect. The results suggest that when many responses are missing, differential nonresponse rates may be a source of DIF.

La plupart des mises en œuvre de la méthode du fonctionnement différentiel des items de MantelHaenszel suppriment les observations ayant des réponses qui manquent ou bien elles remplacent les réponses manquantes par un 0. Ces façons de traiter l'information manquante repose sur des hypothèses quant aux causes des lacunes, hypothèses qui pourraient s'avérer particulièrement inquiétantes quand l'absence de réponse chez les différents groupes suit un schéma distinct. En nous appuyant tant sur des données réelles d'élèves de l'Ontario ayant passé les versions en anglais et en français de deux évaluations à grande échelle, que sur des données simulées, nous avons comparé les indices de la méthode du fonctionnement différentiel des items de MantelHaenszel produits par l'emploi d'un score total ou d'un score de proportion comme variables appariées et en traitant les données manquantes de trois façons : suppression de toutes les observations comportant au moins une valeur manquante; calcul de la corrélation pour chaque couple de variables à partir des données disponibles; ou considération des données manquantes comme des erreurs. Les résultats portent à croire que lorsqu'il manque plusieurs réponses, les taux différentiels d'absence de réponse peuvent constituer une source de fonctionnement différentiel des items.

Barnabas Emenogu is a senior research coordinator in the Literacy and Numeracy Secretariat, Ministry of Education, Toronto. His work involves the application of multivariate and multilevel statistics in the study of large-scale assessments, large-scale educational reforms, policy analyses, and program evaluation.
Olesya Falenchuk works as a research systems analyst in the Education Commons, a role in which she provides research design and analysis consulting to students and faculty and external clients. Olesya holds a doctorate in measurement and evaluation.
Ruth Childs is an associate professor in the Department of Human Development and Applied Psychology.

If a test item is differentially difficult for diverse groups of examinees—that is, if examinees from diverse groups have variable probabilities of answering the item correctly after controlling for overall ability—the item is said to exhibit differential item functioning (DIF). Such items are typically detected using statistical analyses such as the Mantel-Haenszel (MH) DIF procedure (Holland & Thayer, 1988), which is widely used by large-scale testing programs.

Most implementations of the MH DIF procedure require complete response matrices, that is, they cannot accommodate missing data. For example, computer programs for computing MH DIF statistics such as the *Differential Item Functioning Analysis System* (DIFAS, Penfield, 2003) and EZDIF (Waller, 1998) automatically apply listwise deletion, eliminating all examinees with any missing data. Excluding examinees with missing data from the analyses, however, assumes that the data are missing completely at random. This assumption implies that the examinees with missing data are in every other way comparable to those without missing data. If the time limits affect groups of examinees diversely, this assumption is problematic. For example, examinees for whom the language of the test is a second language may not be able to work as fast as other examinees and so may be more likely to have missing data. Translated tests can also be problematic; for example, the French-language versions of large-scale tests are typically longer than the English-language versions because of differences in language structure and vocabulary. This is particularly important when a test has a time limit, as students taking the French-language version may have to do more reading per problem than their counterparts taking the English-language version.

In DIF analyses that require complete data matrices and students with missing data are not to be eliminated, the only way to avoid losing incomplete cases is to replace the missing responses with imputed values: often scores of 0. These treatments make strong assumptions about the causes of the nonresponse. For example, if nonresponse resulted from lack of knowledge, then replacing missing data with scores of 0 makes sense. However, if examinees did not respond because of lack of time—and speed of response is not part of the construct that the test is intended to measure—then it makes much less sense to assume that if they had attempted the items they omitted, they would have responded incorrectly. Varied test-taking approaches are also problematic; for example, examinees who are less willing to guess should not be assumed to have failed to answer because of lack of knowledge. These assumptions may not apply equally to the groups being compared in the DIF analyses.

There is a third alternative for treating missing data in MH DIF analyses (Emenogu, 2006), which, however, is not available in MH DIF analysis programs: to delete incomplete cases only when computing DIF statistics for the items to which those examinees did not respond (analysiswise deletion). In addition, because using the total number of items correct as the matching criterion in MH DIF analyses effectively treats missing responses as wrong, matching on the proportion of items answered correctly out of those attempted may be an appropriate alternative if the assumption is in doubt that the missing data are related to the construct that the test is intended to measure.

The treatment of missing data in DIF analyses may be particularly important when the groups being compared differ in their patterns of missing data. Whereas some educational cultures encourage students to guess on multiple-choice items and attempt every item that requires an extended response, others encourage students to answer only test items that they feel confident they can answer correctly. These differences in test-taking behavior may result in varied nonresponse rates across countries (Schmidt, Wolfe, & Kifer, 1993), or even subpopulations within countries, and may have an effect on the analyses of test data.

Although numerous studies have examined the use of the MH DIF procedure and factors such as sample size or the ability estimate used as the matching criterion, which may affect the results, none has examined the effect of missing data treatment in MH DIF analyses. The purpose of the study described in this article is to identify the effects of the method of handling missing data on Mantel-Haenszel DIF analysis results, especially when differential nonresponse rates occur in the groups being compared. In Study 1, three missing data treatments (listwise deletion of examinees with any missing data, analysiswise deletion, and scoring missing responses as incorrect) and two matching criteria (total score and proportion score) are compared using the responses of Ontario students taking the French- and English-language versions of the 1995 Trends in International Mathematics and Science Study (TIMSS) and the 2001 School Achievement Indicators Program (SAIP) mathematics assessment. Our choice of the latter dataset is based on earlier research that found differential nonresponse rates on the SAIP assessment of the Ontario students taking the French- and English-language versions (Emenogu & Childs, 2005; Xu, 2005).

Although Study 1 has the advantage of using real data, if there are differences among approaches, it is not possible to determine which approach is most accurate as the real incidence of DIF in the datasets is unknown. Therefore, in Study 2, data are simulated based on the missing data patterns in the SAIP data, but with no DIF between groups. Based on the results of Study 2, it is possible to determine which approach is most accurate. Although the data used in this study are from translated tests, this is not a study of translation DIF, which has been studied extensively elsewhere (Emenogu & Childs, 2005; Gierl & Khaliq, 2001; Sireci & Swaminathan, 1996). Nor does this study consider approaches to imputation, as these introduce additional assumptions about why students did not respond.

Method

Data

Study 1. In Study 1, data from two large-scale assessments were analyzed. The first assessment is the 2001 School Achievement Indicators Program Mathematics Assessment, conducted by the Council of Ministers of Education Canada (CMEC) to assess the mathematics achievement of 13-year-old and 16-year-old students. The 2001 SAIP Mathematics Assessment content subtest consisted of 75 multiple-choice items with four response options and 50 short-answer items; both types of items were scored dichotomously. The assessment was administered in two stages (a placement test and a main assessment). Students were first asked to complete a 15-question placement test, which was

scored immediately. The results were then used to classify the students into three groups, and they continued the assessment at three starting points based on their results on the 15-item placement test. Students who scored 0 to 10 were directed to begin the remaining assessment questions at question 16 (starting point 1). Students who scored 11 to 13 were directed to begin at question 41 (starting point 2). Students who scored 14 or 15 were directed to begin at question 66 (starting point 3). Students were asked to complete carefully as many questions as possible in the remaining two and a half hours; however, each student was expected to complete at least 75 items (45 multiple-choice and 30 short-answer). In this study, the responses of the 313 13-year-old and 443 16-year-old Ontario students assigned to the second starting point were analyzed. Of these 756 students, 452 took the English-language version and 304 took the French-language version. Data for students who began at the second starting point were chosen because more students were assigned to this starting point than to either of the other two.

Earlier research found varying nonresponse rates on the SAIP assessment of the Ontario students taking the French- and English-language versions (Emenogu & Childs, 2005; Xu, 2005). For the Ontario students assigned to the second starting point, Figure 1 displays the percentage of students who attempted each item and who answered each item correctly for the 75 items (numbered 1-75, for convenience) that those students were expected to answer. As can be seen from this figure, ratios of correct responses to attempts are similar in the two groups of students. However, a considerably larger percentage of students taking the French-language version did not respond to items. The difference in nonresponse rates especially increases at the end of the test.

In addition to the SAIP data, data from the 41 multiple-choice mathematics and science literacy items in Booklet 1A of the 1995 TIMSS administered to Ontario grade 12 students were analyzed. The TIMSS dataset consisted of the 759 students who took the English-language version of the test and the 318 students who took the French-language version. These students' responses are summarized in Figure 2, which shows that although the students taking the French-language version did have somewhat lower response rates on most items, the response rates are less markedly different than those for the SAIP.

Study 2. The second study used simulated data based on the responses of the 313 13-year-old Ontario students (207 taking the English-language version and 106 the French-language version) who started the SAIP Mathematics Assessment from the second starting point. As differential nonresponse rates were a primary interest for this study, the simulated data were based on the last 25 of the 75 items examined in Study 1 because these items had the largest difference in nonresponse rates between the two groups. The real dataset used as a model for the simulated data, therefore, consisted of 25 dichotomously scored items (15 multiple-choice and 10 short-answer) administered to the 313 13-year-old Ontario students. For these items, 67.1% of English-language examinees and 86.8% of French-language examinees did not respond to one or more test items. Of these 25 items, the average number to which each student did not respond was 3.7 for the English-language group and 8.0 for French-language group.

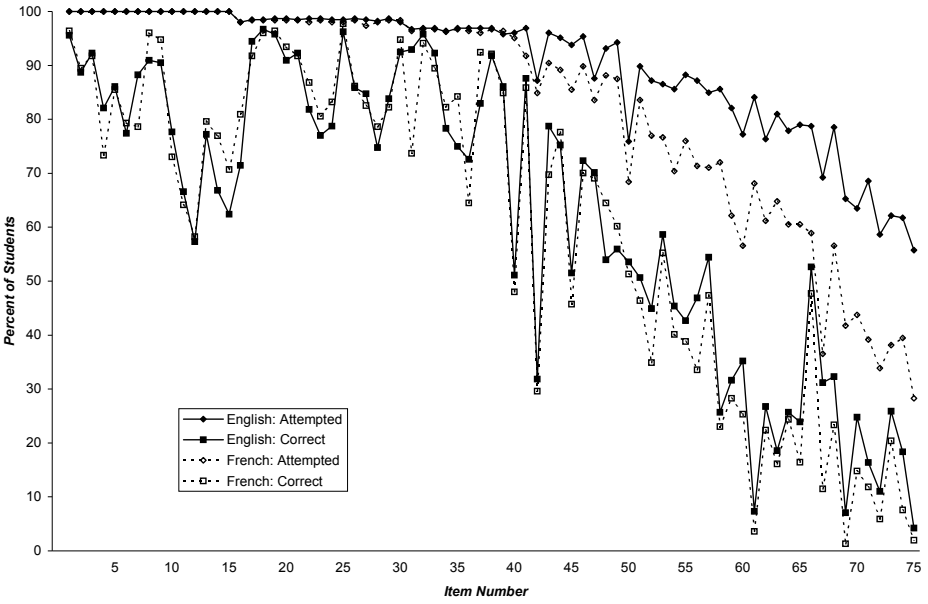

Figure 1. Percentages of students attempting and answering correctly the English- and French-language versions of the 2001 SAIP Mathematics Assessment starting point 2 items.

To obtain item parameters for the data simulation, the real data were calibrated with the two-parameter logistic (2PL) item response theory (IRT) model. Item parameter estimates from this calibration and 4,000 simulees drawn from a normal distribution $N(0,1)$ were used for generation of a complete response matrix. The sample was then randomly split into two samples of 2,000 simulees each. These samples were arbitrarily attributed to the French-language and English-language groups of simulated examinees. By following this procedure, we ensured that the ability distributions of the two groups were the same and that no true DIF was present in the data. Analyses of these simulated data matrices confirmed that none of the items exhibited DIF.

To introduce nonresponse into the generated response matrices for each of the groups, the examinees from the real dataset were classified into 10 ability intervals based on their IRT ability estimates. A contingency table of the proportion of nonresponse for each item and each ability group was created for each of the language groups. These contingency tables were used to generate nonresponse in the simulated data following the non-parametric approach developed by De Ayala, Plake, Impara, and Kozmicky (2000). The distribution of simulated abilities for French- and English-language examinees was divided into 10 segments that were approximately equivalent to the 10 real ability groups. The proportion of nonresponse for each item, contingent on ability group, was compared with a random number. If this was less than or equal to the relative proportion of nonresponse conditional on ability group, then the response was changed to an omission; otherwise, the response remained unchanged. Nonresponse was introduced separately to the samples of French- and English-language simulees based on the proportions of nonresponse in the real data.

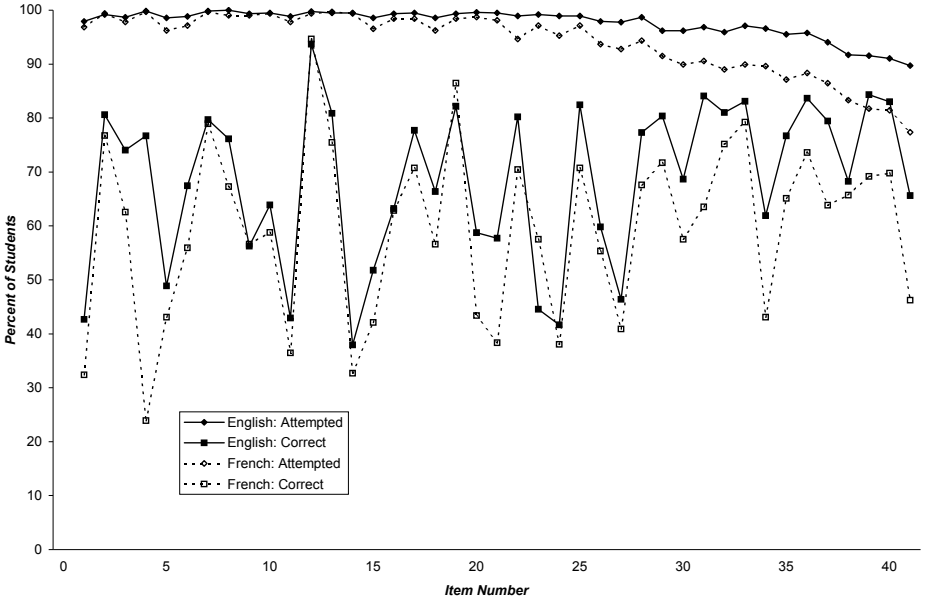

Figure 2. Percentages of students attempting and answering correctly the English- and French-language versions of the 1995 TIMSS Booklet 1A multiple-choice items.

To verify that the simulated data had the properties required for this study, the proportion of simulees attempting each of the 25 items and the proportion of simulees giving correct responses to the attempted items were examined for the reference (English-language) and focal (French-language) groups. The patterns were similar to those for the real responses to the last 25 items (Items 51 to 75) shown in Figure 1.

Analyses

Parallel analyses were performed for the real and simulated data. Combination of the two matching criteria and three nonresponse treatments resulted in six conditions for DIF analysis (Table 1). Both the total score (TS) and proportion score (PS) were divided into 10 matching score categories of equal width. In the listwise deletion (LD) nonresponse treatment, all observations with any missing responses were excluded from the analyses. In missing as incorrect (MI), observations with missing responses were retained but the missing responses were replaced by 0. In analysiswise deletion (AD), cases with missing responses are deleted only when computing DIF statistics for those items that they did not attempt.

For each of the six conditions, the *MH D-DIF* index of differential item functioning was computed as $MH\ D\text{-}DIF = -2.35\ln(\hat{\alpha}_{MH})$, where $\hat{\alpha}_{MH}$ is the estimate of Mantel-Haenszel odds ratio. The standard error for the *MH D-DIF* index was computed with the Robin, Breslow, and Greenland (1987) formula described by Dorans and Holland (1993) and was used to determine the critical values of the *MH D-DIF* at the .05 significance level for each item in each condition. The categories of DIF were specified using the practice of the Educational Testing Service (Dorans & Holland). An item was classified as exhibiting negligible DIF when the *MH D-DIF* index was not significantly different from

Table 1
Conditions for the Computation of MH D-DIF

		Treatment of Missing Data		
		Listwise Deletion (LD)	*All Cases Missing as Incorrect (MI)*	*Analysiswise Deletion (AD)*
Computation of Matching Criterion	Total Score (TS)	TS-LD	TS-MI	TS-AD
	Proportion	PS-LD	PS-MI	PS-AD

zero or its absolute value was less than 1.0. An item was considered to exhibit slight to moderate DIF when the *MH D-DIF* index was significantly different from zero and its absolute value was at least 1.0, and either (a) less than 1.5 or (b) not significantly greater than 1.0. An item was flagged as exhibiting moderate to large DIF when the *MH D-DIF* index was significantly greater than 1.0 and its absolute value was greater than or equal to 1.5.

The *MH D-DIF* analyses were performed using a program written in SPSS (Emenogu, 2006). The matching criterion, whether the total score or proportion score, was not purified because the focus was on the comparison of initial *MH D-DIF* statistics across items.

Results

Study 1: SAIP and TIMSS Data

Of the 75 items analyzed from the 2001 SAIP Mathematics Assessment, three exhibited at least moderate DIF across all six conditions; all these favored the students taking the English-language version of the test. An additional seven were flagged in all conditions except the listwise deletion conditions; three of these favored the English-language version, and four favored the French-language version. Finally, 25 items were flagged for at least moderate DIF in between one and three of the conditions.

There was less nonresponse in the TIMSS data than in the SAIP data and, therefore, greater stability in flagging items for DIF among the six conditions. Five items were identified as having at least moderate DIF in all six treatment conditions, three in favor of the students taking the English-language version of the test and two in favor of the students taking the French-language version. An additional 10 items were flagged between one and four conditions.

The listwise conditions (TS-LD and PS-LD) identified the same items irrespective of how the matching variable was formed; however, how the matching variable was formed led to slightly different results when crossed with the other two missing data treatments. For two items, the magnitude of *MH D-DIF* depended on the matching method; and for another two items, it depended on how missing data were treated, but not on how the matching variable was formed. However, the direction of DIF was consistent. Treating missing responses as incorrect with the proportion score (PS-MI) led to more items being flagged for DIF in favor of students taking the English-language version of the test.

It is important to note that the number of examinees used for the DIF analyses with the diverse treatments varied widely. All examinees were used for computation of *MH D-DIF* when nonresponse was scored as incorrect. Analysiswise deletion resulted in reduction of the sample size used for *MH D-DIF* computation for each item. However, for the listwise deletion, the sample size was reduced dramatically. For the SAIP data, only 108 examinees who took the English-language version and 21 who took the French-language version remained; for the TIMSS, 580 who took the English-language version and 192 who took the French-language version remained. As a result, it was not possible to calculate the *MH D-DIF* statistic for some of the items in the listwise deletion condition because all the remaining examinees in one of the groups answered the item identically (correctly or incorrectly). The listwise deletion approach may not be appropriate when a high proportion of examinees have not responded to one or more items.

Table 2 shows the correlations among the *MH D-DIF* values obtained under the six conditions. Correlations of *MH D-DIF* values for the SAIP dataset are presented below the diagonal; correlations of *MH D-DIF* values for the TIMSS dataset are presented above the diagonal. All coefficients are positive and significant at $p<.01$. The correlations between methods are much higher for the TIMSS data than for the SAIP data. For the SAIP data, the lowest correlation is .59, between the TS-MI and TS-LD conditions, whereas all the coefficients for the TIMSS data are above .93. This might be explained by the fact that the SAIP items have higher proportions of nonresponse than the TIMSS items. The largest coefficients are those between TS-LD and PS-LD for the TIMSS data and between TS-AD and PS-AD for both the TIMSS and the SAIP data, suggesting that how the matching variable is formed has little effect when analysiswise deletion is used.

Study 2: Simulated Data

Analyses of the simulated data using listwise deletion (TS-LD and PS-LD) did not identify DIF in any item for either matching criterion. However, as in Study 1, the LD conditions resulted in dramatically reduced sample sizes: 193 in the English group and 30 in the French group.

Table 2
Intercorrelations of MH D-DIF for the SAIP and TIMSS Data

	TS-LD	TS-MI	TS-AD	PS-LD	PS-MI	PS-AD
TS-LD		.586**	.595**	.996**	.591**	.597**
TS-MI	.959**		.981**	.603**	.946**	.920**
TS-AD	.967**	.989**		.613**	.915**	.968**
PS-LD	1.000**	.959**	.967**		.602**	.613**
PS-MI	.936**	.985**	.957**	.936**		.883**
PS-AD	.963**	.984**	.998**	.963**	.951**	

Note. Intercorrelations for SAIP Mathematics Assessment starting point 2 items (75 items) are presented above the diagonal, and intercorrelations for TIMSS Booklet 1A multiple-choice items (41 items) are presented below the diagonal.
*$p<.05$; **$p<.01$.

When missing responses were scored as incorrect, nine and 12 items were flagged as having at least slight-to-moderate DIF when the total score (TS-MI) and proportion score (PS-MI) respectively were used as the matching criterion. Of the nine items identified under TS-MI, five favored the reference group and six favored the focal group; all 12 items identified in the PS-MI condition favored the reference group. Seven of these items were flagged as exhibiting moderate-to-large DIF. No common items were flagged for both matching criteria.

Analysiswise deletion of nonresponse produced three items favoring the focal group with total score as a matching criterion (TS-AD). Two of these items (Items 53 and 54) were the same as those flagged for moderate DIF when nonresponse was scored as incorrect. Analysiswise deletion with proportion correct scoring (PS-AD) did not falsely identify DIF in any items.

Table 3 shows the correlations among the *MH D-DIF* values obtained for the simulated data under the six conditions for treatment of nonresponse, plus the dataset with no nonresponse. The *MH D-DIF* statistics obtained by treating nonresponse with listwise deletion (TS-LD and PS-LD) are not significantly correlated with those from the other nonresponse treatments. The correlations between *MH D-DIF* statistics when nonresponse was scored as incorrect (TS-MI and PS-MI) and treated with analysiswise deletion (TS-AD and PS-AD) are moderate and significant. Most important, only the TS-AD and PS-AD results are significantly correlated with those obtained from the complete response matrix, r=.78, p<.01 and r=.83, p<.01; the difference between the correlations is not significant, however, Williams's $T^2(22)$=–0.81, p=.42 (Steiger, 1980). These results suggest that analysiswise deletion produces the most valid *MH D-DIF* statistics.

Discussion

The purpose of this study was to examine the effect of the method of handling missing data with the occurrence of differential nonresponse rates on the results of Mantel-Haenszel DIF analyses. The results suggest that when there are few missing responses, as in the TIMSS dataset, the choice of missing data treatment will not matter much. However, when there are many missing

Table 3
Intercorrelations of MH D-DIF Statistics for the Simulated Data

	No Nonresponse	TS-LD	TS-MI	TS-AD	PS-LD	PS-MI	PS-AD
TS-LD	–.06						
TS-MI	.33	–.25					
TS-AD	.78**	.04	.08				
PS-LD	–.10	1.00**	.08	.03			
PS-MI	.12	–.28	1.00**	.03	–.31		
PS-AD	.83**	.14	.41*	.86**	.11	.12	

Note. Correlations based on simulated data for 25 items.
*p<.05; **p<.01.

responses, differential nonresponse rates in the reference and focal groups may indeed serve as a source of DIF. In particular, analyses of the simulated data shows that scoring of nonresponse as incorrect falsely identifies items as exhibiting DIF whether the total score or proportion correct of the attempted items is used as the matching criterion. This suggests that scoring nonresponse as incorrect is not the best option of nonresponse treatment in *MH DIF* analyses, especially in the presence of differential nonresponse. Analysiswise deletion also falsely identified some items as exhibiting DIF when the total score is used as the matching criterion. However, with proportion correct as the matching criterion, this problem was not evident. Listwise deletion also did not flag any items as exhibiting DIF; however, all students that miss at least one item are removed from all analysis with this nonresponse treatment, which results in great reduction of the sample size for computation of the *MH D-DIF* statistics and of the power of the analyses to detect DIF.

In summary, these results suggest that the best approach when there is differential nonresponse is to use analysiswise deletion for the treatment of nonresponse and proportion correct of attempted items as the matching variable. The generalizability of these results are, of course, limited by the use of only two real datasets and a simulation with a single rate of nonresponse and by the fact that DIF analyses may be influenced by data traits. Further investigation with a wider range of patterns of nonresponse is needed. Also needed are investigations of the reasons for varied nonresponse rates across populations.

This study nevertheless adds substantially to our knowledge about the effects of missing data treatment on *MH DIF* analyses and highlights the importance of carefully considering the assumptions implicit in selection of missing data treatments, especially when investigating DIF between groups such as students in Ontario's French- and English-language schools that differ in their patterns of nonresponse. In such a study, of course, care must be used in applying listwise deletion as this may lead to increased Type II error rates.

References

De Ayala, R.J., Plake, B.S., Impara, J.C., & Kozmicky, M. (2000, April). *The effect of omitted responses on ability estimation in IRT*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Emenogu, B. (2006). *The effect of missing data treatment on Mantel-Haenszel DIF detection*. Unpublished doctoral dissertation, Ontario Institute for Studies in Education of the University of Toronto.

Emenogu, B., & Childs, R. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education, 28,* 128-146

Gierl, M.J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164-187.

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

Penfield, R.D. (2003). *DIFAS 1.1: Differential Item Functioning Analysis System users manual*. Miami, FL: University of Miami.

Schmidt, W.H., Wolfe, R.G., & Kifer, E. (1993). The identification and description of student growth in mathematics. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom processes* (pp. 59-99). Oxford, UK: Pergamon.

Sireci, S.G., & Swaminathan, H. (1996, October). *Evaluating translation equivalence: So what's the big DIF?* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Waller, N.G. (1998). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and Logistic Regression procedures. *Applied Psychological Measurement, 22*, 391.

Xu, Y. (2005). *Examining group differences in omission rates in a mathematics assessment*. Unpublished master's thesis, Ontario Institute for the Studies in Education of the University of Toronto.