Practical Assessment, Research, and Evaluation

Volume 13 *Volume 13, 2008* Article 7

2008

An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating

Michalis P. Michaelides

Recommended Citation

Michaelides, Michalis P. (2008) "An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating," *Practical Assessment, Research, and Evaluation*: Vol. 13 , Article 7.
DOI: https://doi.org/10.7275/n04d-8767
Available at: https://scholarworks.umass.edu/pare/vol13/iss1/7

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 13, Number 7, September 2008 ISSN 1531-7714

An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating

Michalis P. Michaelides, *European University Cyprus*

In this study the Mantel-Haenszel procedure, widely used in studies for identifying differential item functioning, is proposed as an alternative to the delta-plot method and applied in a test-equating context for flagging common items that behave differentially across cohorts of examinees. The Mantel-Haenszel procedure has the advantage of conditioning on ability when making comparisons of performance of two examinee groups on an item. There are schemes for interpreting the effect size of differential performance, which can inform the decision as to whether to retain those items in the common-item pool, or to discard them. Data from a statewide assessment are analyzed to illustrate the use of this procedure. Advantages of this methodology are discussed and limitations regarding test design that may make its application difficult are described.

Test equating methods are statistical adjustments that establish comparability between alternate forms built to the same content and statistical specifications by placing scores on a common scale (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). In the common-item nonequivalent groups design for test equating (Kolen & Brennan, 2004), a subset of the items is embedded in two (or more) test forms to provide a common basis for comparing examinee groups that respond to different forms. The information obtained from the common items serves to attribute any differences in test performance to group ability differences and/or to test form differences.

When an equating procedure is performed, the underlying assumption is that the set of anchor items works the same way with both groups (Wainer, 1999). For the common-item nonequivalent groups design to provide valid equating results, the common items must function similarly in both forms (Hanson & Feinstein, 1997). If two groups of examinees respond differently to an item, then it might not be appropriate to include that item in the equating process. The problem of inconsistent behavior of common items across administrations can be viewed as an instance of differential item functioning (DIF), where two groups taking two different forms with some items in common are the focal and reference groups.

In practice, a procedure for examining the volatility of equating items' difficulty values is the delta-plot method, a simple and comprehensible way for studying the item-by-group interaction, which makes use of the item p-values (Angoff, 1972). The delta-plot is a graphical procedure that flags outliers in a plot of the transformed p-values of the common items. It is widely implemented because it is practical, does not entail time-consuming calibrations such as those required in an Item Response Theory (IRT) framework, and provides prima-facie evidence regarding anomalous changes in item difficulties across administrations. However, it is a crude procedure in the sense that it summarizes the information from an item in a single number, the p-value, and looks at how that number is related to the p-values of the remaining common items. Moreover, it transforms the p-values through an inverse normal transformation, which changes their distribution in a somewhat arbitrary way.

In this study an alternative to the delta-plot procedure is implemented: the Mantel-Haenszel statistic (Mantel & Haenszel, 1959) is applied to detect differential item performance. The Mantel-Haenszel statistic is commonly used in studies of DIF, because it

makes meaningful comparisons of item performance for different groups, by comparing examinees of similar proficiency levels, instead of comparing overall group performance on an item. Overall group performance versus performance stratified by proficiency could result in dissimilar outcomes, a statistical paradox known as Simpson's (1951) paradox: a group U may have a higher proportion correct on an item than a group V; however after dividing the distributions of the two groups into strata (e.g. on the basis of proficiency level), group-V subgroups may have higher proportion correct indices than their matched group-U subgroups. In essence, the overall p-values would imply that group U is at an advantage, as would a delta-plot analysis, while it would be at a disadvantage according to the stratified analysis and the results of a procedure such as the Mantel-Haenszel statistic (for a numeric example see Dorans & Holland, 1993).

THEORETICAL FRAMEWORK

Consistent behavior is a desirable characteristic that common items are expected to have when administered to different groups. Particularly within the context of IRT, the property of parameter invariance promises that item parameter estimates remain relatively unchanged across various groups of examinees and ability estimates remain invariant across groups of items (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). If the IRT model fits the data perfectly, then parameters will be invariant across administrations, except for sampling fluctuations that introduce random error in the responses of examinees. In that case, the changes in the behavior of item parameter estimates would follow a systematic pattern depending on the changes in the size and proficiency of the different examinee groups.

IRT makes strong assumptions and its promise for invariance depends on the degree that the model assumptions, and particularly unidimensionality, hold (Miller & Linn, 1988). Many studies have concluded that items do not always behave in consistent ways; item indices and IRT item parameter estimates of the same items differ when obtained from different administrations. A common explanation has been content effects, such as discrepancies in instructional coverage (Miller & Linn, 1988; Yen, Green, & Burket, 1987), opportunity to learn (Masters, 1988), changes in curricular or instructional emphasis (Bock, Muraki, & Pfeiffenberger, 1988; Sykes & Fitzpatrick, 1992). A second type of explanation for differential item difficulty for different groups is context effects such as speededness, fatigue, test wiseness (Masters, 1988), changes in the positioning of the item (Kingston & Dorans, 1984; Yen, 1980), the time lapse between testing and the classroom teaching of the content (Cook, Eignor, & Taft, 1988), and disclosure of, or familiarity with items (Mitzel, Weber, & Sykes, 1999).

The abundance of evidence demonstrating the possibility of differential item behavior across forms raises the question of how to deal with these items particularly when they belong in the common-item group, a case that is not unusual (e.g. Michaelides, 2003). Misbehaving common items may be removed from the anchor group. However, changes in item behavior may indicate a true change in the proficiency of the examinee population. Indeed, the educational community may have reallocated resources in the system to achieve changes and these are properly reflected in differential item behavior. Consequently, automatic deletion of differentially behaving common items may compromise the validity of the testing program.

The delta-plot method

In the delta-plot procedure, p_{Yi}, the proportion correct of a common item i in administration Y (here $Y=1,2$ stands for two administrations, for example in Year-1 and Year-2, that have some items in common) is transformed to the delta metric, δ_{Yi}, through a linear transformation of the inverse normal equivalent (Dorans and Holland, 1993) as follows:

$$\delta_{Yi} = 13 - 4\,\Phi^{-1}(p_{Yi}) \qquad (1)$$

In the delta metric, a p-value of 0.5 is transformed to 13, larger delta values correspond to more difficult items and smaller delta values to easier items, as opposed to the proportion correct scale, which is bounded between 0 and 1 with easier items having higher values than more difficult ones.

When two groups respond to the same items, the item p-values, p_{Yi}, for each group are calculated, transformed to the delta metric with equation 1, and plotted on a scatterplot. Each point corresponds to an item with a delta value, δ_{1i}, for the Year-1 group plotted on the horizontal axis and a delta value, δ_{2i}, for the Year-2 group plotted on the vertical axis. The (δ_{1i}, δ_{2i}) points should form a straight line pattern. If the items are equally difficult in the two administrations, then the

points should fall on, or, due to sampling error, roughly on the identity line. Outliers from the general trend of the plot denote items that are functioning differentially for the two groups with respect to the level of difficulty.

A handy rule to determine which items are outliers is to draw a "best-fit" line to the points and calculate the perpendicular distances of each point to the line. The fitted line is chosen to minimize the sum of squared perpendicular distances (not the sum of squared vertical distances as in ordinary least squares regression) of the points to the line. This is known as "principal components" or "principal axis" regression and unlike ordinary least squares regression, it is symmetric: the line obtained by regressing the independent on the dependent variable and the line obtained by regressing the dependent on the independent variable are identical. The straight line is fitted as shown in equation 2.

$$y = \frac{s(\delta_{2i})}{s(\delta_{1i})} x + \bar{\delta}_{2i} - \frac{s(\delta_{2i})}{s(\delta_{1i})} \bar{\delta}_{1i} \qquad (2)$$

The distance of each point to the best-fit line is then calculated. Any points lying more than three standard deviations of all distances away from that line are flagged as outliers. Such items call for inspection to determine plausible causes for the differential performance in the two groups, and are candidates for exclusion from the common item pool. Analysts seek to determine possible reasons for why the items functioned differentially. They can speculate whether the differential performance is related to the purpose of measurement, i.e. if it reflects a true change in the proficiency of the examinee cohorts, or if it is due to superficial or irrelevant conditions, such as a change in the positioning of the item. An item may then be discarded from the equating-item pool and treated as a regular, non-common item. Inclusion or exclusion of an item from the equating pool is a matter of judgment and affects the equating function.

The Mantel-Haenszel statistic

The Mantel-Haenszel common odds ratio assesses the strength of association in three-way *2x2xj* contingency tables. It estimates how stable the association between two factors is in a series of *j* partial tables. Holland and Thayer (1988) published a paper that popularized the Mantel-Haenszel statistic as a potential method for studying DIF in any two groups of examinees. It can be used to test the null hypothesis

$$H_0 : \frac{p_{Ri}}{q_{Ri}} = \frac{p_{Fi}}{q_{Fi}}$$

Namely, the odds for answering correctly item *i* for the reference group *R*, $\frac{p_{Ri}}{q_{Ri}}$, are equal to the corresponding odds for the focal group *F*, $\frac{p_{Fi}}{q_{Fi}}$. Note that $q_{\bullet i} = 1 - p_{\bullet i}$. The alternative hypothesis is

$$H_1 : \frac{p_{Ri}}{q_{Ri}} = \alpha_j \frac{p_{Fi}}{q_{Fi}}$$

where $\alpha_i = \frac{p_{Ri} q_{Fi}}{p_{Fi} q_{Ri}}$ is the common odds ratio ($\alpha_i \neq 1$.)

The focal and reference groups are matched on ability using a test score interval as a proxy. The procedure provides a chi-square test statistic as well as an estimator of α_i across the *j 2x2* tables. The latter is a measure of the effect size, or how much the data depart from H_0, an important feature since conventional statistical significance can be easily obtained with large enough samples.

The Mantel-Haenszel procedure may be implemented in the context of equating with the common-item nonequivalent groups design to identify which common items behave differentially across two forms/administrations by considering the two examinee cohorts as the focal and reference groups.

METHODOLOGY AND DATA SOURCES

The delta-plot and the Mantel-Haenszel methods were implemented using data from a grade 4 statewide Visual and Performing Arts (VPA) assessment. The test consisted of 12 spiraled forms and a total of 84 items. Each form comprises 7 items, 6 of which are dichotomous, scored 0 or 1, and 1 polytomous, scored on a 0-4 scale. Dichotomous items were multiple-choice, and polytomous items were constructed-response questions. It was administered in two consecutive years. Of the 84 items, 69 were common over the two annual administrations and distributed across the forms according to Table 1. Each common item appeared in only one form. About 1300 examinees responded to each form. The last two columns on the table show that the average performance

and standard deviations on the common items between the two student groups who responded to corresponding forms were very similar. For instance, the students who took Form 2 in Year 1 had 6 items in common with those who took Form 2 in Year 2. From those items, 5 were dichotomously scored and one was polytomously scored with a maximum score of 4, therefore the maximum number correct score on the Form 2 common items was 9.

Table 1. *Characteristics of the 12 forms of the assessment*

Form	Total number of common items across years	Number of polytomous common items	Max. no. correct score on common items	Mean (and sd of) no. correct score on common items	
				Year 1	Year 2
1	5	1	8	4.1 (1.64)	3.8 (1.61)
2	6	1	9	4.7 (1.78)	4.7 (1.70)
3	7	1	10	5.7 (1.89)	5.5 (1.74)
4	6	1	9	4.6 (1.81)	4.2 (1.71)
5	5	-	5	3.4 (1.33)	3.4 (1.35)
6	7	1	10	6.0 (1.83)	6.1 (1.79)
7	5	1	8	4.4 (1.71)	4.3 (1.69)
8	6	-	6	4.0 (1.41)	4.1 (1.38)
9	6	1	9	5.1 (1.63)	5.0 (1.57)
10	5	1	8	4.6 (1.65)	4.3 (1.58)
11	6	1	9	4.5 (1.83)	4.5 (1.72)
12	5	-	5	2.3 (1.35)	2.5 (1.34)
Total	69	9			

First, the delta-plot procedure was carried out. The p-values for the dichotomous items and the mean score over maximum score for the polytomous items in each of the two annual administrations were transformed to the delta metric by equation 1. A point for each item was plotted on a scatterplot using its two delta values, one from each administration. A principal axis regression line was fitted to all the points. Any point that lay more than three standard deviations of all distances of the points to the line away from the line was labeled as an outlier, a common item that behaved differentially across the two years.

The Mantel-Haenszel procedure was applied next. The examinees taking Form X in Year 1 and those taking Form X in Year 2 constituted the reference and the focal group respectively. A number-correct score for each examinee was derived by summing his/her scores on the common items. For example, the examinees taking Form 9 had number-correct scores ranging from 0 to 9, because there were 5 dichotomous and 1 polytomous common item in that particular form. The number-correct score serves as the matching criterion *j*.

For each dichotomous common item *i*, a *2x2xj* three-dimensional table was constructed. One variable was the group each examinee belonged to (Year-1 or Year-2) and the second was his/her score on the item (0 or 1). The matching variable was the third dimension j of that table, *j=0,…, K*, where *K* is the maximum number-correct score on the common items of the form.

Table 2. *Tabulation of counts for the j^{th} partial table for a dichotomous common item i*

		Score on item *i*		
		1	0	Total
Group	Year 1	A_j	B_j	N_{1j}
	Year 2	C_j	D_j	N_{2j}
	Total	M_{1j}	M_{0j}	T_j

As can be seen in Table 2, for one such partial table *j* of the dichotomous common item *i*, the counts of the correct (*Aj* or *Cj*) and incorrect (*Bj* or *Dj*) responses for each of the two examinee cohorts were recorded. *Tj* stands for the total count of responses for item *i* on the j^{th} partial table.

Using the counts from Table 2, the estimate of the Mantel-Haenszel common odds ratio, $\hat{\theta}_{MH}$ for a common item *i*, is given by

$$\hat{\theta}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}$$

The common odds ratio takes values from 0 to infinity; a value of one implies that there is no differential item performance between two groups, and larger values imply that the item favors the reference group.

The Educational Testing Service (ETS) has developed a scheme for classifying items into categories of DIF that considers both statistical significance and magnitude of the log-odds ratio. The log-odds ratio is transformed to the delta metric, and referred to as the Mantel-Haenszel delta difference (MH D–DIF; Dorans & Holland, 1993), by

$$MH\ D-DIF = -2.35\ ln\left(\hat{\theta}_{MH}\right)$$

If for an item MH D-DIF is equal to 1.0, then that item was easier for the focal than for the reference group by one delta point. Formulae for calculating the variability of the common-odds ratio and the MH D-DIF can be found in Appendix 1.

A dichotomous item is classified into one of three categories: A, B, and C, which correspond to negligible, intermediate, and large DIF. The classification rules (Dorans & Holland, 1993; Zieky, 1993) are as follows:

- **Category A** if the MH D-DIF is not significantly different from zero ($p \geq 0.05$), or if its absolute value is less than 1.0.
- **Category B** if the MH D-DIF is significantly different from zero, its absolute value is at least 1.0, and its absolute value is either less than 1.5 or not significantly greater than 1.0.
- **Category C** if the MH D-DIF is significantly greater than 1.0 in absolute value, and its absolute value is at least 1.5.

There are extensions of the Mantel-Haenszel procedure for cases where the levels of a variable are more than two. In addition to their *2x2xj* analysis, Mantel and Haenszel (1959) proposed a generalized statistic for more than two response categories in a variable. A chi-square test for the case of more than two ordered response categories was provided by Mantel (1963). Scores need to be assigned to each category, and a deviation of the sum of cross products from the expectation, and its variance conditioned on all marginal totals can be computed. Table 3 demonstrates how the data can be arranged in general and in the case of a *0-T* scored polytomous item *i* on a j^{th} partial table.

Table 3. *Counts for the j^{th} partial table for a polytomous common item i*

		Score on item *i*				
		$Y_1=0$	$Y_2=1$	…	$Y_{T+1}=T$	Total
Group	Year 1	N_{10j}	N_{11j}	…	N_{1Tj}	N_{1+j}
	Year 2	N_{20j}	N_{21j}	…	N_{2Tj}	N_{2+j}
	Total	N_{+0j}	N_{+1j}	…	N_{+Tj}	N_{++j}

According to Mantel (1963) the chi-square statistic under the null hypothesis of no association is

$$Mantel`s\ \chi^2 = \frac{\left(\sum_j \sum_T N_{2Tj} Y_T - \sum_j \frac{N_{2+j}}{N_{++j}} \sum_T N_{+Tj} Y_T\right)^2}{\sum_j Var\left(\sum_T N_{2Tj} Y_T\right)} \quad (3)$$

for a polytomously scored item *i*, with *j* partial tables. The terms are given by the partial tables, as shown on Table 3. Under the null hypothesis of the common odds equal to one, *Mantel's* χ^2 has a chi-square distribution with one degree of freedom. For the purposes of differential item behavior, rejecting the null hypothesis suggests that members of two subpopulations matched on a measure of proficiency differ in their mean performance on the item under investigation (Zwick, Donoghue, & Grima, 1993).

As in the case of the dichotomous items, judgments as to whether a polytomous item exhibits DIF or not take into account a measure of effect size in addition to statistical significance. Dorans and colleagues (Dorans & Kulick, 1986; Dorans & Schmitt, 1991/1993) proposed a measure of the standardized mean differences, which

compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations. Zwick, et al. (1993) reformulated the Standardized Mean Difference (SMD) as follows:

$$SMD = \sum_j \frac{N_{2+j}}{N_{2++}} \frac{\sum_T N_{2Tj} Y_T}{N_{2+j}} - \sum_j \frac{N_{2+j}}{N_{2++}} \frac{\sum_T N_{1Tj} Y_T}{N_{1+j}} \qquad (4)$$

The first term in the SMD is the mean performance on the item for the Year-2 group. Subtracted from that is the mean item performance for the Year-1 group weighted by the Year-2 group distribution of the matching criterion. In Appendix 1, the variance formula for the SMD is presented.

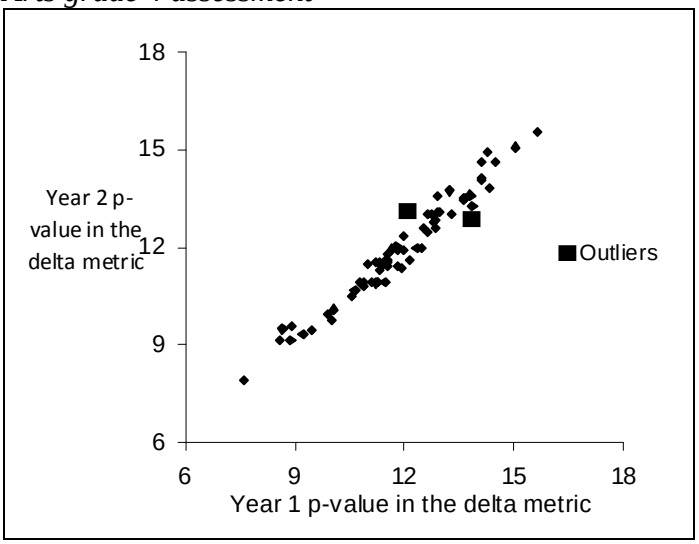
One approach to analyzing polytomous items would be to dichotomize them by choosing a cut-point on the scoring scale and assigning a correct response to the scores above the cut, and an incorrect for the scores below. Then they can be treated as dichotomous with the same ETS scheme described above[1]. Since statistics have been developed to deal with DIF with polytomous items, similar empirical rules exist to guide decisions on whether a polytomous item exhibits DIF or not. The rules combine statistical significance given through Mantel's chi-square statistic (equation 3) and a measure for the magnitude of the difference between the performances of the two groups. The effect size is the SMD (equation 4) divided by the within-group standard deviation of the studied item, pooled over the two groups. A generalization of the scheme for the dichotomous items is used in the National Assessment of Educational Progress (NAEP) to classify the polytomous items for DIF (U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2001). The rules for category assignment are:

- A polytomous item is classified in **Category AA** if either Mantel's chi-square is not significantly different from zero $(p \geq 0.05)$, or if the absolute value of the effect size is less than or equal to 0.17.
- A polytomous item is classified in **Category BB** if Mantel's chi-square is significant and the absolute value of the effect size is over 0.17 and less than or equal to 0.25.
- A polytomous item is classified in **Category CC** if Mantel's chi-square is significant and the absolute value of the effect size is over 0.25 (J. Donoghue, personal communication, June 17, 2003).

RESULTS

The delta-plot of the common items in the Visual and Performing Arts grade 4 assessment appears in Figure 1. The delta-plot procedure flagged two dichotomous items as outliers: 4.1 and 7.1. The former was the first common item in form 4 and was easier for the Year-1 cohort (p-value=0.59, delta-value=12.07 compared to 0.49 and 13.14 respectively for the Year-2 cohort.) The latter was the first common item in form 7 was easier for the Year-2 cohort (p-value=0.51, delta-value=12.90 compared to 0.42 and 13.82 respectively for the Year-1 cohort.)

Figure 1. *Delta-plot for the Visual and Performing Arts grade 4 assessment*



[1] The polytomous items were initially treated as dichotomous, after the scores were dichotomized to "0" for scores 0 and 1 and "1" for scores 2, 3, and 4. A different dichotomization was examined with "0" for scores 0, 1, and 2 and "1" for scores 3 and 4. These dichotomizations are both arbitrary. The former gave p-values for the dichotomized polytomous items that were closer to the difficulty values of the polytomous items than the latter. Under this treatment, a common odds ratio, log-odds and the variance of the log-odds can be computed for each polytomous item and entered into the scheme for deciding whether to flag an item for DIF or not. The results of the dichotomized polytomous common items are presented in Michaelides (2003). In this paper, the polytomous items are not dichotomized, but are analyzed with the appropriate statistics (Mantel's chi-square and SMD).

With the delta-plot method, the decision to flag an item as an outlier is confounded with the differences in the shape of the ability distributions of the two examinee groups. The Mantel-Haenszel procedure circumvents this problem by comparing the item performance of examinees with similar proficiency scores, thus adjusting for differences in the shapes of the ability distributions of the two groups. For each common item, the relevant statistics were calculated: odds and log-odds ratio, and the standard deviation of the log-odds for the dichotomous items; for each polytomous item a table with the multiple ordered response categories with the associated SMD, its standard deviation, and Mantel's chi-square statistic. Applying the Mantel-Haenszel procedure, three items were flagged for intermediate DIF. Their statistics appear on Table 4. The dichotomous item 5.1 and the polytomous item 7.5 favored the Year-1 cohort. Item 7.1, which was identified by the delta-plot procedure as well, favored the Year-2 cohort.

Table 4. *Common items flagged for DIF by the Mantel-Haenszel procedure*

Item	Item type	MH D-DIF	Standard Error (MH D-DIF)	ETS Category
5.1	Dichotomous	-1.2380	0.3131	B
7.1	Dichotomous	1.3175	0.2147	B

Item	Item type	SMD (pooled SD)	Effect size	Mantel CHI-SQ	ETS Category
7.5	Polytomous	-0.1716 (0.9489)	-0.1809	53.06	BB

Departures from unidimensionality could arise if items are measuring different skills and could result in flagging more items for DIF (Welch & Miller, 1995). To investigate the dimensionality of the data, principal components analysis of the scores on the common items within each form was conducted. In 19 of the 24 cases (12 forms by 2 years) only one principal component was extracted with an eigenvalue larger than one. Two principal components were extracted in 4 forms that included a polytomous item and in one form with dichotomous items only. In those cases, the loadings of items on the two principal components did not differentiate between the types of items. Histograms of the number-correct score on a form showed that the distributions of the matching criterion between Year 1 and Year 2 are very similar (Michaelides, 2003).

Another concern is the number of total counts in partial tables. In the forms 5 and 7, where outliers were flagged, total counts in partial tables were large enough (minimum total count $T_{j=0} = 33$ in form 7). There were some instances in the analysis of other forms where the total count for a partial table was smaller; the minimum that had occurred was 13. Results when total counts are small should be interpreted more cautiously, and adjacent partial tables could be collapsed to avoid this problem.

The two procedures result in findings that are not always in agreement. Perpendicular distances of each point from the fitted line on the delta-plot are graphed against an effect size that involves the Mantel-Haenzel D-DIF for the dichotomous items (Figure 2a) or the SMD for the polytomous items (Figure 2b). In both graphs there is a positive association between the variables; the correlation coefficients are 0.54 for the dichotomous items and 0.60 for the polytomous items, a moderately high relationship illustrating that the two methods produce quite similar indices for flagging items that behave differentially. One common item, 7.1, was identified by both procedures since it was more than three standard deviations of all distances (3 x 0.166) away from the delta-plot fitted line, and was classified as category B on the ETS classification scheme that relies on the Mantel-Haenszel statistic. There are differences in the flagged items between the two procedures: item 4.1 which was flagged due to its large distance from the delta-plot fitted line, was not identified by the Mantel-Haenszel method due to its low MH D-DIF. Item 5.1 had a large MH D-DIFF to be placed in category B, but it was within three standard deviations of the distances from the delta plot line. Finally, 7.5, the only polytomous item that was flagged as BB by the Mantel-Haenszel procedure because of an effect size

(absolute value) slightly larger than the 0.17 cut-point was not flagged by the delta-plot method. The analysis of this dataset indicates that despite the positive association between the calculated measures from the two methods, there are substantial differences among them.

Figure 2a. *Comparison of the two methods – dichotomous items*

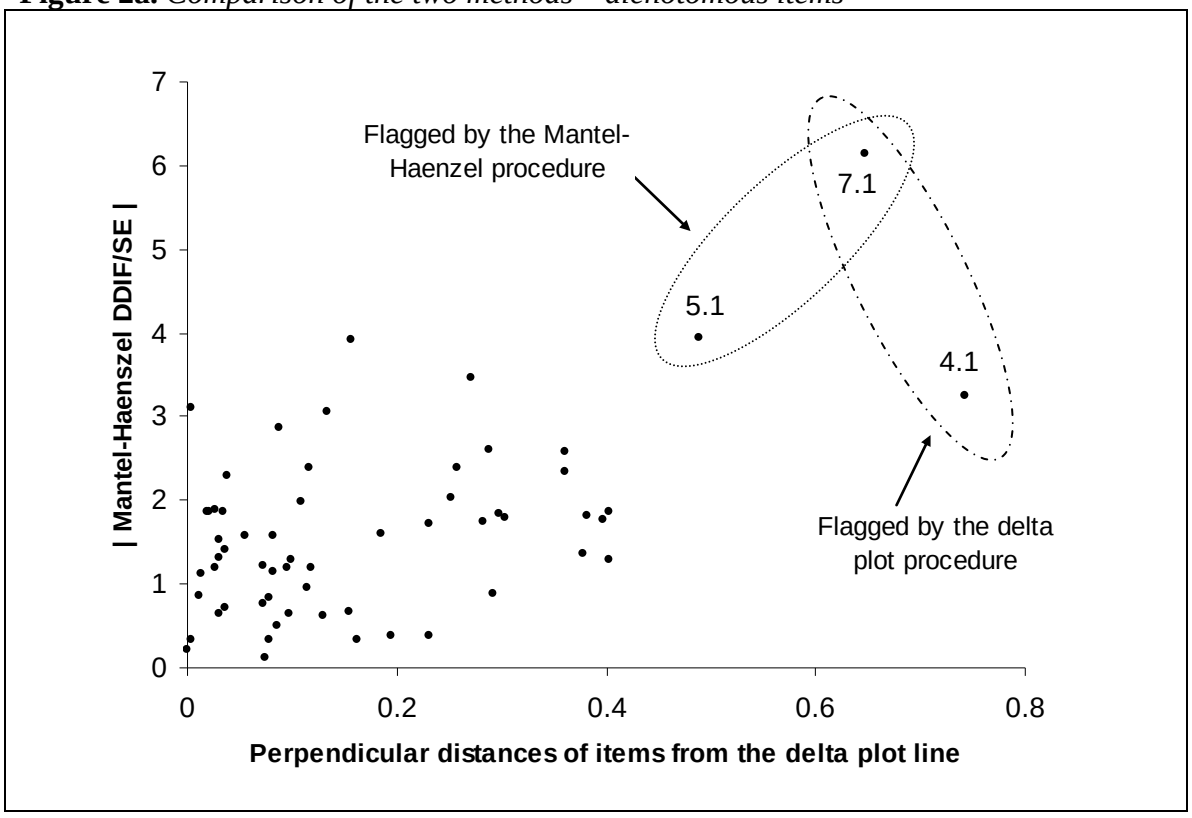
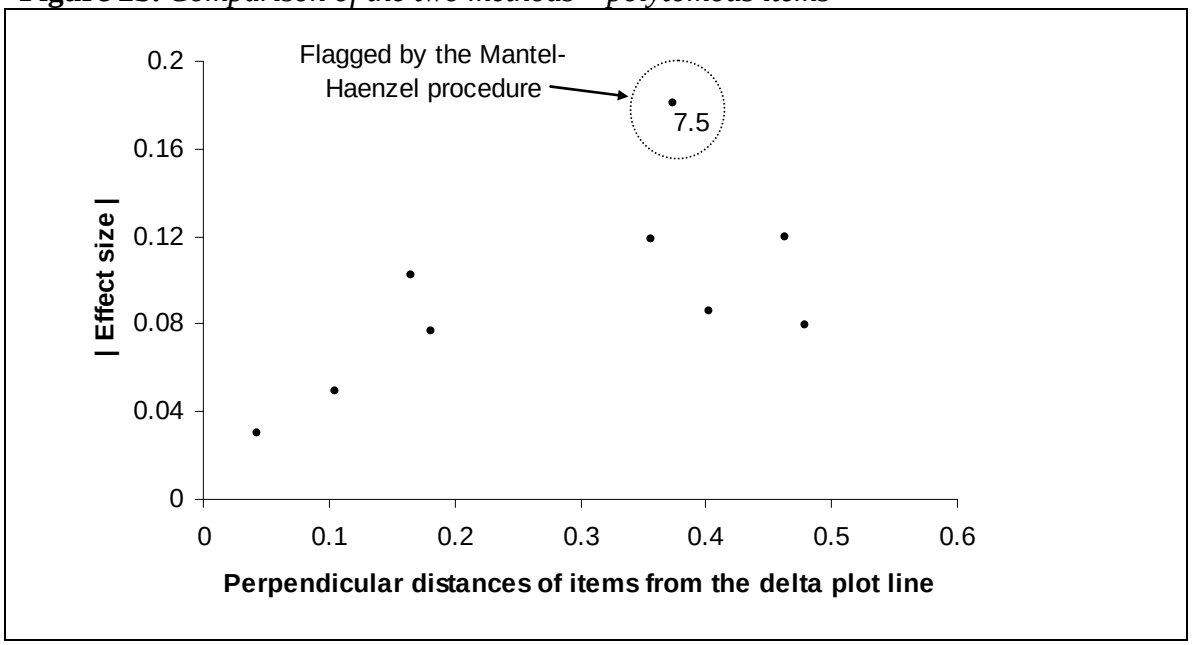


Figure 2b. *Comparison of the two methods – polytomous items*



Some remarks on the matching criterion

The performance of comparable members of the two groups is contrasted to detect differential item behavior. Holland and Thayer (1988) define comparability as "identity in those measured characteristics in which examinees may differ and that are strongly related to performance on the studied item" (p.130). In the test analyzed, all examinees who took a form in Year 1 are compared to the examinees who took the corresponding form in Year 2. Corresponding forms have a number of common items embedded in them. Matching is done based on the number correct score on a set of common items administered to both groups of examinees, summing all item scores, without rescaling the scores on the polytomous items (Zwick, et al., 1993). The number-correct score is the usual choice in studies of DIF (Welch & Miller, 1995).

Because of the use of a score on a test in which the studied item appears and which includes the score on the studied item, the Mantel-Haenszel procedure involves some circular reasoning in what it purports to do: evaluation of differential performance on an item that taps a construct after controlling for a proxy for the performance on a domain that includes the same construct. However, the choice of a total test score may be the best available matching criterion because it is a common measure that exists for all examinees; it is typically reliable as long as the test is validated for its intended purposes; and it is more reliable than individual items (Dorans & Holland, 1993). In the current study the score on the common items was used as a matching criterion since that was the longest, common measure that both Year-1 and Year-2 subgroups had taken.

Holland and Thayer (1988) conjectured that when an item is analyzed for DIF it should be included in the matching criterion, but if it exhibits substantial DIF it should be excluded when examining other items. They showed that, under the Rasch model when the studied item is included in the matching score the null hypothesis for the Mantel-Haenszel holds in the population; when the studied item is excluded and there is no DIF, the procedure does not behave correctly. Zwick (1990) concurred with Holland and Thayer's findings and argued that inclusion of the studied item improves the behavior of the odds ratio with more general models as well. Donoghue, Holland, and Thayer (1993) showed that inclusion of the studied item in the matching variable results in reducing the number of false positives, i.e. items that do not behave differentially being flagged as exhibiting non-zero DIF. In general, there is agreement that the studied item must be included in the matching criterion, although when there are polytomous items in a test, their inclusion or exclusion will have a larger effect on the criterion (S-H Kim, Cohen, Alagoz, & S. Kim, 2007)[2].

An underlying assumption with the use of the Mantel-Haenszel procedure for the study of DIF is that the items studied are homogeneous and unidimensional (Angoff, 1993). Unidimensionality is more than an assumption for studies of DIF; it is a part of the definition of item bias (Shepard, 1982). An item that functions differentially for a group is an item that measures a construct that departs from what the matching criterion measures. If it did not, then performance of the group on that item would not be expected to depart from that predicted by the group's performance on the overall criterion.

While items that are scored on a scale with multiple points, such as essay prompts or performance assessments, are considered to capture important aspects of student knowledge that are difficult to assess with more traditional testing formats, they are likely to introduce additional dimensions unrelated to the measured construct. These other dimensions may be sources of differential group performance. The dimensionality of the matching is always a concern; it is perhaps more crucial when there are polytomous items involved. Because of the more complicated nature of open-ended performance tasks, Zwick, et al. (1993) state that construct-irrelevant factors could interfere with the intended construct and lead to larger differences between groups. Nonetheless, they generalize from Holland and Thayer's (1988) dichotomous case that polytomous items should be included in the matching variable by simply summing the scores on all, dichotomous and polytomous, items. For practical reasons, this could be the only option, since tests that include polytomously scored tasks tend to have fewer

[2] Additional analyses where conducted in which the matching criterion was *purified* (Dorans & Holland, 1993), i.e. the items that were flagged with intermediate DIF were excluded from the calculation of the number correct score on the common items. The same items, and no other common items, were flagged in the same DIF category under the refined criterion (Michaelides, 2005).

items (Welch & Miller, 1995). If a matching criterion consists of very few items, the reliability of the stratification on ability will be low. Finding an appropriate matching criterion may be difficult (Dorans & Schmitt, 1991/1993), in fact impossible, if the some of the available items are not used because they are polytomous.

DISCUSSION

When two groups do not perform equally well on an item, then the item exhibits differential impact with respect to the two groups. Impact is often stable and could replicate in other similar items, since overall ability attributes will contribute to this disparity. When the differences in ability distributions are accounted for by matching the groups on a relevant characteristic, if there are still differences between the similar subgroups, these are unexpected given the similarity of the groups on an attribute that the item and the matching ability proxy are supposed to measure (Dorans & Holland, 1993). Matching on a relevant third variable and then comparing what is comparable has become a central concern in the study of DIF; it is crucial in making the distinction between differences in item p-values attributable to differences in item functioning versus differences in group ability (Dorans & Schmitt, 1991/1993).

Conditioning on a criterion is common to most methods of studying DIF. The delta-plot method, which was originally proposed as a technique for detecting and studying item-by-group interactions (Angoff, 1972) takes into account changes in the mean and standard deviation of the item difficulties by fitting a best-fit line. It disregards however further information about differences in the distributions of ability, and thereby confounds group differences in ability distributions with group differences in how examinees of a given ability find an item. The more the shapes of two ability distributions differ, the more the confounding is amplified when two single numbers, the p-values, are compared. It is now considered to be technically flawed for examining item bias (Angoff, 1993).

Hence, the Mantel-Haenszel procedure which relies on raw scores, as well as logistic regression approaches and IRT-based methods which rely on IRT calibrations that are more complicated and time-consuming, have taken over in studies of DIF (see Camilli (2006) for a classification of DIF analysis methods). If two administrations of a test form with common items embedded in both are considered as subgroups in a DIF-like study, then the Mantel-Haenszel procedure could provide more refined comparisons between examinees that are similar, and flag items that exhibit either homogeneous odds greater than or less than 1 across all levels of ability, or certain differential patterns of odds across ability levels.

The content of the assessment analyzed herein could raise concerns about dimensionality. A test assessing skills on topics such as visual arts, music, and theater for children in grade 4 includes questions that address quite different content, proficiency, or skill, especially when polytomous items are added to the common-item pool. The subjectivity involved in scoring items for creativity, cultural understanding, and aesthetics could also result in inconsistent scoring. The scorers of the Year-1 and the Year-2 administrations could be quite different in their scoring patterns, thus introducing additional dimensions in the scores.

For the particular assessment the Mantel-Haenszel procedure seemed to work well. More items were flagged as exhibiting DIF than outliers identified by the delta-plot, while one of the items was flagged by both methods. Principal component analysis and histograms of ability distributions did not raise serious concerns about departures from dimensionality that could result in detecting false-positive occurrences of DIF. The idiosyncratic features of the Visual and Performing Arts assessment might have been expected to lead to many items being flagged for DIF. However, very few items were flagged. With more common items embedded in corresponding forms the stratification of the ability variable could be more refined, matching subgroups that were even more similar in ability, increasing sensitivity of the Mantel-Haenszel procedure to actual DIF versus false-positives.

The delta-plot did not identify any polytomous items. The Mantel-Haenszel procedure flagged one. Item 7.5 had a SMD of –0.1716 which suggests that there is some moderate difference in the performance (or scoring) of the two cohorts. As with all flagged items, the next step would be to inspect it through analysis of the content to detect whether it is unique compared to other items in the matching criterion. Scoring analysis of polytomous items can reveal whether the observed differential performance was due to inconsistent scoring and not due to actual examinee responses. The Year-2 scorers could re-score randomly selected responses of

Year-1 examinees and compare their scoring practice with the Year-1 scores to detect any differential patterns.

Test design and implementation issues

There are certain requirements on the test design that need to hold to apply the Mantel-Haenszel procedure for the study of differential behavior of common items across administrations. The Visual and Performing Arts grade 4 assessment analyzed in this study was the only test out of more than twenty statewide assessments inspected that did not violate requirements of a Mantel-Haenszel implementation. All assessments were constructed under a matrix-sampling design; there were multiple test forms in each of two annual administrations. A form in the Year-2 administration corresponded to a form in the Year-1 administration because they shared a set of common items. However, for purposes of linking the Year-2 to the Year-1 assessment, it is not two corresponding forms that are equated, but all alternate forms of one year to all alternate forms of the other year. What happened in most available data sets was that a common item appearing in a form X in Year 1 would be moved to a different form Y in Year 2. The common-item pools were not identical across administrations, a practice that should be avoided given the research findings on context effects reviewed in the theoretical framework section of this paper (see also Michaelides, 2003 for a review). In other cases, the number of alternate forms from one year to the next changed and to conduct equating some of the common items were rearranged in newly introduced forms. Yet a few data sets that avoided moving common items around forms had as few as three or four common items in corresponding forms. Some writing assessments had only polytomous items as common. The assessment that was eventually analyzed was the only one that for all forms had a proper matching criterion, i.e. for all forms, all the common items of a Year-1 form appeared in the corresponding Year-2 form, and which consisted of a number of items that was not prohibitively low.

In DIF studies these problems are not likely to emerge because the groups usually compared, gender or ethnicity groups, take the same test form, thus all test items, not just the common items, can be part of the matching criterion. In the case of equating and using Mantel-Haenszel statistics to study the behavior of the common items, design issues are more complicated. The matching criterion can only be as long as the common-item pool, i.e. a fraction of the total test. Provided that the size of the common-item pool is large enough, the matching criterion could be refined enough to provide many strata for reliable matching. But if the common items are spread across many forms, as is the case of matrix-sampling designs, and especially if there are polytomous items in the assessment, the number of items used to form the matching variable is limited.

CONCLUSIONS

The choice of items to include as common in a common-item nonequivalent groups design influences the equated scores and their accuracy. In this study, the Mantel-Haenszel procedure was proposed as an alternative to the delta-plot method and applied in the context of test equating for flagging common items that behaved differentially across cohorts of examinees. The Mantel-Haenszel procedure has the advantage of conditioning on ability when making comparisons of performance of two groups on an item. There are schemes for interpreting the effect size of differential performance, which can inform the decision as to whether to retain those items in the common-item pool, or to discard them. However, there may be some test-design limitations that preclude the application of this procedure in a test-equating framework.

It is not easy to provide strict guidelines on how to deal with common items flagged for differential behavior across two forms. The content tested by a common item and its relevance to both the curriculum framework and actual instruction comes into the decision as to how to treat it, if it behaves in unexpected ways. As in the case of DIF studies where an instance of an item functioning differentially for two groups does not necessarily imply that the item is biased and should be discarded from a test (Linn, 1993), finding a common item that fails to function consistently across administrations does not imply that it is inappropriate for equating. If a common item is flagged by the Mantel-Haenszel procedure (or the delta-plot method) as behaving differentially in any two forms, it does not mean that it should be automatically discarded from the common item pool and treated as a new, non-common item in the second form. Content experts and test developers may be able to offer plausible explanations for the differential behavior. If a context effect has, for example, been discovered, then it is probably legitimate to say that it is unrelated to the construct that the test is measuring. However, as regards equating, even in obvious cases of discrepant performance due to irrelevant circumstances, discarding a common item is not as straightforward. Common

items are chosen to meet certain content and statistical specifications, and to proportionally represent the properties of the total test. Discarding a common item might violate those guidelines and introduce a different kind of bias in the equating transformation.

Even though this judgmental step is involved in the equating process, the practice can be improved in different ways. For example, the impact of including or excluding an item on the specifications and the content representation of the common-item pool can be examined. If exclusion of an item violates those specifications seriously, e.g. if it is the single item from a particular area of the tested subject, then discarding it may not be advisable. Another instructive piece of information is the effect that a single common item can have on the equating transformation and the equated scores. With knowledge of a common item's leverage, the decision on how to deal with it can be more informed. A third way would be the kind of information given by the schemes for flagging items for DIF implemented in this study. Inclusion of the effect size of the differential behavior, in addition to the statistical significance, to characterize the amount of DIF and labeling it as negligible, intermediate, or large, can be useful in deciding whether a flagged item should be discarded or not.

Limitations and further research

Beyond the extent of influence that misbehaving common items can have on equating results, the reasons behind the unexpected behavior are worthy of investigation. The content and the context of the common items were not examined herein, although they have some bearing on the decision as to how to deal with the outliers, as has been discussed.

The uniqueness of each testing program and the specific situations under which items are administered make it difficult to devise preset rules for dealing with misbehaving common items. As with studies of DIF, numbers by themselves cannot provide definite decision rules with regard to complicated and sensitive issues of DIF and fairness (Zieky, 1993). Equating practice can be augmented, however, by more informative procedures. How to apply a Mantel-Haenszel procedure to flag items with a real data set has been empirically demonstrated. Characterizing the effect size of the differential behavior of items further facilitates judgments as to how to treat them. An additional useful tool would be a procedure for evaluating the influence of each item, given its characteristics: type of item, position on a scatterplot, distance from a best-fit line, etc. Studies that simulate realistic situations would provide insight on the importance of item characteristics that affect the leverage of outliers. Vukmirovic, Hu, and Turner (2003) ran one such simulation, but defined outliers on a scatterplot of IRT *b* parameters instead of p-values on a delta-plot. Using such information together with the plausible causes of differential item behavior – content, context, or unidentifiable – the decision to keep or discard a common item can be more defensible.

Moreover, there are additional DIF procedures that have been developed, for instance methods based on logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999) that provide effect size measures. These methodologies could be adapted in the test equating context for examining common items. The effectiveness of each method is currently an issue being investigated in the literature (e.g. Kim et al., 2007; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Su & Wang, 2005).

In conclusion, the decision on how to treat misbehaving common items is judgmental, and we have shown in previous work that including or excluding as few as one or two delta-plot outliers might impact equated score aggregates by a substantial amount (Michaelides, 2003). The Mantel-Haenszel procedure can be more informative than the delta-plot because performance on items is conditioned on a measure of proficiency, thus flagging items that function differentially for similar examinees. Schemes that classify items for DIF taking into account both the magnitude and the statistical significance of the log-odds can be implemented to determine the amount of DIF for each common item and inform the decision whether to include or exclude any flagged items from the common-item pool. Although on theoretical grounds it is argued that the Mantel-Haenszel procedure is superior to the delta-plot in identifying differential item behavior, there are some design issues that may preclude its application in an equating context. Maintaining a consistent matching criterion across two forms is necessary to carry out the procedure, and a large number of common items must exist for the matching criterion to be long enough to make reliable categorizations of examinees. These two conditions may not hold, especially under matrix-sampling assessment designs where the common items are spread across different forms.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Angoff, W. H. (1972, Sept.). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)

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

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Camilli, G. (2006). Test Fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th edition, pp. 221-256). Westport, CT: ACE/Praeger.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of irt and conventional item parameter estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the mantel-haenszel and standardization measures of differential item functioning. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service. [Also appears in *Construction vs. Choice in Cognitive Measurement* (pp. 135-166), R. E. Bennett & W. C. Ward (Eds.), 1993, Hillsdale, NJ: Erlbaum.]

Hanson, B. A., & Feinstein, Z. S. (1997). *Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item Equating Design* (ACT Research Report Series No. 97-1). Iowa City, IA: ACT Inc.

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, S.-H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF Detection and Effect Size Measures for Polytomously Scored Items. *Journal of Educational Measurement, 44*(2), 93-116.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating Scaling, and Linking: Methods and Practices* (2nd edition). New York: Springer.

Kristjansson, E., Aylesworth, R, McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mantel, N. (1963). Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

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.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

Michaelides, M. P. (2003). *Effects of common-item selection on item response theory test equating with nonequivalent groups*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Michaelides, M. P. (2005). *An application of a Mantel-Haenszel procedure to flag misbehaving common items in test equating*. Poster presented at the 2005 Annual Meeting of the American Educational Research Association, Montreal, Canada.

Miller, A. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Mitzel, H. C., Weber, M. M., & Sykes, R. C. (1999). *Test item disclosure: How much difference does it really make?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore: John Hopkins University Press.

Simpson, E. H. (1951). Interpretation of interaction contingency tables. *Journal of the Royal Statistical Society, (Series B), 13*, 238-241.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*, 313-350.

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

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT *b* values. *Journal of Educational Measurement, 29*(3), 201-211.

U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. *The NAEP 1998 Technical Report*, NCES 2001-509, by Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). Washington, DC: National Center for Education Statistics.

Vukmirovic, Z., Hu, H., & Turner, J. C. (2003, April). The effects of outliers on IRT equating with fixed common item parameters. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il.

Wainer, H. (1999). Comparing the Incomparable: An Essay on the Importance of Big Assumptions and Scant Evidence. *Educational Measurement: Issues and Practice, 18*(4), 10-16.

Welch, C. J., & Miller, T. R. (1995). assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32*(2), 163-178.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*(4), 297-311.

Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid information from customized achievement tests. *Educational Measurement: Issues and Practice, 6*(1), 7-13.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187-201.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321-344.

.

APPENDIX 1.

Variance or standard error calculations for the statistics used in the paper

Variability statistics of the common odds ratio and the MH D-DIF:

The natural logarithm of the common odds ratio $\hat{\theta}_{MH}$, the "log-odds" has a symmetric distribution centered at zero. Holland and Thayer (1988) report the following approximation of the variance of the log-odds derived by Phillips and Holland (1987)

$$Var\left[ln\left(\hat{\theta}_{MH}\right)\right] = \frac{1}{2\left(\sum_j A_j D_j / T_j\right)^2} \sum_j \left(A_j D_j + \hat{\theta}_{MH} B_j C_j\right)\left[A_j + D_j + \hat{\theta}_{MH}\left(B_j + C_j\right)\right] / T_j^{\,2}$$

The quotient of the log-odds ratio with its standard deviation can be compared to the standard normal distribution for statistical significance.

The MH D-DIF has a standard error (Dorans & Holland, 1993) of

$$SE(MH\ D-DIF) = 2.35\sqrt{Var[ln(\hat{\theta}_{MH})]}$$

Variability of Mantel's chi-square statistic and of the Standardized Mean Difference:

The variance terms in the denominator of Mantel's chi-square (equation 3) are

$$Var\left(\sum_T N_{2Tj} Y_T\right) = \frac{N_{1+j} N_{2+j}}{N_{++j}^2 \left(N_{++j} - 1\right)}\left[N_{++j} \sum_T N_{+Tj} Y_T^2 - \left(\sum_T N_{+Tj} Y_T\right)^2\right]$$

Zwick and Thayer (1996) provided a standard error for the SMD based on Mantel's (1963) multivariate hypergeometric model and one based on a two-multinomial model. The former performed better in their simulation study. In a comparative study by Zwick, Thayer, and Mazzeo (1997), the SMD as a descriptive index performed best among three descriptive statistics of polytomous item DIF and together with the former standard error, as good as other 5 inferential methods when the two subpopulations had the same distribution. The hypergeometric variance of the SMD is reported in this paper and calculated as follows:

$$Var(SMD) = \sum_j \left(\frac{N_{2+j}}{N_{2++}}\right)^2 \left(\frac{1}{N_{2+j}} + \frac{1}{N_{1+j}}\right)^2 Var_H\left(\sum_T N_{2Tj} Y_T\right)$$

The variance terms are defined as those in the denominator of Mantel's chi-square (equation 3).

Acknowledgement

This research was partially supported by CRESST/UCLA, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education. Their support is gratefully acknowledged. Ideas and opinions stated do not necessarily represent CRESST/UCLA positions or policies. The author thanks Edward Haertel for his constructive guidance and insightful ideas in conducting this study; Measured Progress for the use of their data; participants at presentations of this research at Stanford University, the Educational Testing Service, and CTB/McGraw-Hill, as well as anonymous reviewers for their comments. An earlier version of this paper was presented at the 2005 Annual Meeting of the American Educational Research Association.

Citation

Michaelides, Michalis P. (2008). An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating. *Practical Assessment Research & Evaluation*, 13(7). Available online: http://pareonline.net/getvn.asp?v=13&n=7

Author

Michalis P. Michaelides
Assistant Professor, Educational Psychology
European University Cyprus

Email: M.Michaelides [at] euc.ac.cy