

On the Study of Differential Item Performance without IRT

Paul W. Holland
Educational Testing Service
Princeton, New Jersey 08541

1. INTRODUCTION

The problem of identifying items for which the performance of certain subpopulations -- often women and minorities -- is unusual and out of line with their performance on other items or test results has a substantial history. The book by Berk (1982) summarizes the state of the art as of 1980 and the work of Lord (1980), Scheuneman (1979), Shepard, et.al. (1981), among others are relevant. From a statistical point of view, modern methods for the analysis of multi-way contingency tables seem particularly appropriate to this problem and some suggestions for their use have been made, (Marascuillo and Slaughter, 1981). In this spirit, the present paper proposes the well-known method of Mantel and Haenszel (1959) for the analysis of $2 \times 2 \times K$ contingency tables as an easily implemented, powerful technique for the measurement of the degree to which two subpopulations of examinees perform differently on a given test item. Modern references to the Mantel-Haenszel procedure include Breslow (1981), Hauck (1979), and Breslow and Liang (1982). The basis for the use of the Mantel-Haenszel (herein MH) procedure in the study of differential item performance is the fundamental notion of the need to compare comparable people when examining the relative performance of two groups of examinees on an item. This is the problem of matching and is discussed in section 2. Section 3 gives the relevant facts about the MH procedure while section 4 discusses various aspects of the MH procedure and related methods in the context of measuring differential item performance.

2. MATCHING VERSUS CONTROLLING FOR ABILITY

The need to "control for ability" is well established in the differential item performance literature. It is the fundamental basis for the proposed use of item response theory methods to study "item bias." Other methods, such as those of Scheuneman (1979), use test performance as a proxy for ability. The "delta-plot method," Angoff (1982), controls for ability indirectly by concentrating attention on the covariance between the item difficulty indices for the two groups rather than on their respective mean values.

In my opinion, the "need to control for ability" is an inadequate way to express a more fundamental idea. When we compare two subpopulations on any criterion, it is always important to be sure that only comparable members of the two groups are being compared. What constitutes comparability will depend on the problem at hand. In the study of differential item performance we are interested in learning something about a test item and how members of one subgroup (the "focal group") might react differently to it than do the members of another subgroup (the "reference group"). If our criterion is performance (i.e., right or wrong on the test item) then it is improper to compare the performance of reference and focal group members who differ in significant and

measurable ways that are related to their performance on the item. Differential item performance means differences in performance on an item between focal and reference group members that is attributable to characteristics of the item and not to differences in characteristics of the groups of examinees.

When we confound both examinee characteristics and item characteristics and simply look at differences in the performance on an item of reference and focal group members we are measuring what is called impact rather than differential item performance. For example, comparing the proportion of reference and focal group members who give correct answers to a given item is a measure of the item's impact on the focal group relative to the reference group. In measuring differential item performance members of the reference and focal groups are first divided into sets of examinees who are matched on relevant criteria before their performance on the item is compared. Examples of relevant matching criteria are: scores on related tests, schooling measures, and other group membership. In many practical settings, matching will be done on related test scores since these are both available and accurately measured.

The 2x2xK Table: For a given item, say item j , the data from the i^{th} matched group of reference and focal group members can be arranged as a 2x2 table:

	Right on item j	Wrong on item j	
Reference	A_i	B_i	n_{Ri}
Focal	C_i	D_i	n_{Fi}
Total	R_i	W_i	n_{+i}

(1)

For $i=1, \dots, K = \text{number of matched groups}$. In (1) A_i denotes the number of reference group members in the i^{th} matched group who answered item j correctly. B_i , C_i , and D_i have corresponding interpretations. n_{Ri} and n_{Fi} denote the number of reference and focal group members, respectively, in the i^{th} matched group, while n_{+i} denotes the total number in the i^{th} matched group of examinees. R_i and W_i denote the number in the i^{th} matched group who get the item right and wrong, respectively. Considered together these K 2x2 tables form one big 2x2xK table. There is one such 2x2xK table for each item being considered. It is worth emphasizing that once the criteria for matching have been selected, the 2x2xK table of data can be formed from samples of data from the reference and focal group members. It should also be emphasized that the choice of matching variables is important and will depend on the availability, amount, and accuracy of data as well as on its relevance to item performance.

3. THE MANTEL-HAENSZEL PROCEDURE

In the i^{th} matched group, the odds that a reference group member gets item j correct is A_i/B_i , while the corresponding odds for a focal group member is C_i/D_i . The MH procedure measures the advantage (or disadvantage) on item j that reference group members have relative to their matched focal group colleagues by the ratio of these two odds. This gives us the odds-ratio estimate

$$\hat{\alpha}_i = \frac{A_i}{B_i} / \frac{C_i}{D_i} = \frac{A_i D_i}{B_i C_i} . \quad (2)$$

The $\hat{\alpha}_i$ estimate a population cross-product-(or odds-) ratio, α_i , for the i^{th} matched group.

The Mantel-Haenszel common-odds-ratio estimate is a weighted average of the $\hat{\alpha}_i$ that uses the following weighted formula:

$$\hat{\alpha}_{MH} = \frac{\sum \omega_i \hat{\alpha}_i}{\sum \omega_i} , \quad (3)$$

where

$$\omega_i = \frac{B_i C_i}{n_{+i}} . \quad (4)$$

Substituting (4) into (3) yields the usual formula for $\hat{\alpha}_{MH}$:

$$\hat{\alpha}_{MH} = \frac{\sum A_i D_i / n_{+i}}{\sum B_i C_i / n_{+i}} . \quad (5)$$

The Mantel-Haenszel estimate, $\hat{\alpha}_{MH}$, is the average factor by which the likelihood that a reference group member gets item j correct exceeds the corresponding likelihood for comparable focal group members. (Likelihood is measured by the odds of getting item j correct). For example, if $\hat{\alpha}_{MH} = 1$ then reference and focal group members are, averaging across all the matched groups, equally likely to be correct on the item. When $\hat{\alpha}_{MH} > 1$ then the reference group has the advantage whereas when $\hat{\alpha}_{MH} < 1$ the focal group has the advantage.

Associated with the estimate $\hat{\alpha}_{MH}$ is a one-degree-of-freedom chi-square test of the hypothesis that all of the population cross-product ratios in all of the 2×2 layers of the $2 \times 2 \times K$ table are unity (i.e., $\alpha_i = 1$ all i). This test is given by the formula:

$$\chi_{MH}^2 = \frac{(|\sum_i A_i - \sum_i \mu_i| - \frac{1}{2})^2}{\sum_i \sigma_i^2} \quad (6)$$

where

$$\mu_i = E(A_i | \alpha_i=1) = \frac{nR_i R_i}{n_{+i}} \quad (7)$$

and

$$\sigma_i^2 = \text{Var}(A_i | \alpha_i=1) = \frac{nR_i nF_i R_i W_i}{(n_{+i})^2 (n_{+i}-1)} \quad (8)$$

The χ_{MH}^2 from (6) will be large if $\hat{\alpha}_{MH}$ differs from 1.0 significantly in either direction. Thus, this test will detect differential item performance that favors either the reference or the focal group.

4. DISCUSSION

The MH procedure is closely related to log-linear model procedures for estimating a constant two-way interaction across a series of 2x2 tables (see Bishop, Fienberg, and Holland, 1975). In practical terms, $\hat{\alpha}_{MH}$ is usually nearly identical to estimates of the common-odds-ratio that involve complicated iterative procedures. While the formula for $\hat{\alpha}_{MH}$ is a simple weighted average of the sample odds-ratio $\hat{\alpha}_i$, it has been shown (Breslow, 1981) that, over the range of values relevant to this application of the MH procedure, $\hat{\alpha}_{MH}$ is nearly optimal as an estimator. In other words, no other estimate of the common-odds-ratio can have a substantially smaller variance. The chi-square test based on χ_{MH}^2 is of high power because it is concentrated into a single degree of freedom rather than dissipated across several degrees of freedom.

If there is more than one pair of groups that could serve as the reference and focal group in an analysis then values for $\hat{\alpha}_{MH}$ and χ_{MH}^2 can be computed for all such pairings.

The parameter $\Delta = -2.35 \ln(\alpha)$ is (approximately) in the scale of differences in delta-units of difficulty where delta-units are those used by ETS in their normal item analysis procedures. This transformation can be used to put $\hat{\alpha}_{MH}$ values into units which are familiar to those who use the delta-scale in test construction and analysis:

$$\hat{\Delta}_{MH} = -2.35 \ln(\hat{\alpha}_{MH}). \quad (9)$$

Thus, $\hat{\Delta}_{MH} = -1.0$ means that the focal group found the item one delta-unit harder

than did comparable members of the reference group. The parameter $\hat{\Delta}_{MH}$ is similar to an average shift to the right of $-\Delta_{MH}$ in the ICC of the focal group relative to the ICC of the reference group in an IRT model (as estimated by LOGIST).

The MH procedure can easily be expanded to include an analysis of distractor choice for multiple choice tests. For five-choice responses the 2x2 table in (1) is replaced by the following 2x6 table

Response on item j

	A	B	C*	D	E	Omit	Total
Reference							n_{Ri}
Focal							n_{Fi}
Total							n_{+i}

C* is correct answer, for example. Then the MH procedure is applied to the five 2x2 tables formed by juxtaposing the column for the correct answer with a column for one of the five ways of producing incorrect answers. E.g.,

	C*	A		C*	B		C*	Omit
Reference							
Focal								

This yields five MH cross-product estimates and five chi-square tests for each item. In some cases these may be used to see if a significant value of $\hat{\alpha}_{MH}$ is due to a single type of incorrect answer.

There are a number of important research issues that need to be addressed in the use of the MH procedure in the study of differential item performance.

What aspects of the criteria for matching examinees seriously affects $\hat{\alpha}_{MH}$ in practical settings -- the reliability of the criteria, the fineness of the matching, the use of other examinee attributes, etc.? How stable are the values of $\hat{\alpha}_{MH}$ across different examinee populations? What are the relationships between the values of $\hat{\alpha}_{MH}$ and other statistical indices used to construct tests -- i.e., difficulty and discrimination? How should values of $\hat{\alpha}_{MH}$ for several pairs of reference and focal groups be combined for the same test item?

The MH procedure promised to be a relatively inexpensive and yet statistically powerful technique for identifying test questions that are potentially "biased" or unfair in some way to identified groups of examinees. ETS is currently embarked on a variety of research projects to see how to best use this tool for such purposes.

References

- Angoff, W. (1982) "Use of difficulty and discrimination indices for detecting item bias" in Berk, R. (ed.) Handbook of Methods for Detecting Test Bias. Baltimore and London: Johns Hopkins University Press.
- Berk, R. (Ed.) (1982) Handbook of Methods for Detecting Test Bias. Baltimore and London: Johns Hopkins University Press.
- Breslow, N. (1981) Odds ratio estimates when the data are sparse. Biometrika, 68, 73-84.
- Breslow, N. and Liang, K. (1982) The variance of the Mantel-Haenszel estimator. Biometrics, 38, 943-952.
- Hauck, W. (1979) The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics, 35, 817-819.
- Lord, F. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.
- Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.
- Marascuillo, L. and Slaughter, R. (1981) Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.
- Scheuneman, J. (1979) A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.
- Shepard, L., Camilli, G., and Averill, M. (1981) Comparisons of six procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.