Journal of Educational Measurement
Winter 1990, Vol. 27, No. 4, pp. 361–370

Detecting Differential Item Functioning Using Logistic Regression Procedures

Hariharan Swaminathan
University of Massachusetts
and
H. Jane Rogers
Teachers College, Columbia University

A logistic regression model for characterizing differential item functioning (DIF) between two groups is presented. A distinction is drawn between uniform and nonuniform DIF in terms of the parameters of the model. A statistic for testing the hypothesis of no DIF is developed. Through simulation studies, it is shown that the logistic regression procedure is more powerful than the Mantel-Haenszel procedure for detecting nonuniform DIF and as powerful in detecting uniform DIF.

The detection of item bias, or differential item functioning (DIF), in achievement, licensure, and credentialing examinations has become an important issue in recent years. Many methods, ranging from analysis of variance approaches to item response theory techniques, have been proposed (see Shepard, Camilli, and Averill, 1981). Currently, the most promising of these methods appears to be IRT-based methods and the Mantel-Haenszel procedure proposed by Holland and Thayer (1988). The Mantel-Haenszel procedure is particularly attractive because it is easy to implement and has an associated test of significance.

Simple and practical though it may be, the Mantel-Haenszel procedure is not designed for and may not be powerful in detecting nonuniform DIF. Uniform and nonuniform DIF have been defined by Mellenberg (1982). *Uniform* DIF exists when there is no interaction between ability level and group membership. That is, the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability. *Nonuniform* DIF exists when there is interaction between ability level and group membership: that is, the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels. In item response theory terms, nonuniform DIF is indicated by nonparallel item characteristic curves.

The occurrence of nonuniform DIF has been documented by Mellenberg (1982) and Hambleton and Rogers (1989). Mellenberg found that a Word Analogies Test given to Tanzanian and Kenyan students contained several items showing nonuniform DIF. Hambleton and Rogers (1989) found that a number of items in the 1982 New Mexico High School Proficiency Examination showed nonuniform DIF when Anglo- and Native Americans were compared. These

The authors would like to thank Paul Holland and Wendy Yen for their comments and suggestions on an earlier version of this article.

authors, using an IRT-based method and the Mantel-Haenszel procedure, found that the Mantel-Haenszel statistic was unable to detect this type of DIF.

Mellenberg (1982) used the log-linear model to predict item responses from group membership, ability level, and the interaction of these factors. A nonzero interaction term indicates the presence of nonuniform DIF. This procedure is analogous to a factorial analysis of variance procedure where ability levels are considered discrete unordered categories. A drawback of Mellenberg's procedure is that it ignores the ordered nature of the ability levels and hence does not use all of the available information. A method which treats ability as a continuous variable can be expected to be more powerful.

Item response theory procedures (Shepard et al., 1981; Hambleton & Swaminathan, 1985) do take into account the continuous nature of ability when comparing the performance of groups of examinees. The drawback of IRT-based procedures is that they are sensitive to sample size and model-data fit and are expensive in terms of computer time. In addition, indexes such as the area between item characteristic curves have no associated tests of significance.

The purpose of this paper is to provide a procedure that extends both the Mantel-Haenszel and the Mellenberg procedures and that provides a cost-effective alternative to IRT-based methods. This procedure, based on the logistic regression model, takes into account the continuous nature of the ability scale and is capable of identifying both uniform and nonuniform DIF. Models of this nature have been used by Bennet, Rock, and Kaplan (1987) and Spray and Carlson (1986) to study differences between groups on dichotomous item responses.

The Logistic Regression Procedure

The logistic regression model for predicting the probability of a correct response to an item is

$$P(u = 1|\theta) = \frac{e^{(\beta_0+\beta_1\theta)}}{[1 + e^{(\beta_0+\beta_1\theta)}]},$$

where u is the response to the item, θ is the observed ability of an individual, β_0 is the intercept parameter, and β_1 is the slope parameter. This is the standard logistic regression model for predicting a dichotomous dependent variable from given independent variables (Bock, 1975).

The logistic regression model given above can be used to model differential item functioning by specifying separate equations for the two groups of interest:

$$P(u_{ij} = 1|\theta_{ij}) = \frac{e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}}{[1 + e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}]}, \quad i = 1, \ldots, n_j, j = 1, 2. \qquad (1)$$

Here u_{ij} is the response of person i in group j to the item, β_{0j} is the intercept parameter, β_{1j} is the slope parameter for group j, and θ_{ij} is the ability of individual i in group j. The accepted definition of DIF is that an item shows DIF if

individuals with the same ability but from different groups do not have the same probability of success on the item (Hambleton & Swaminathan, 1985). It follows that no DIF is present if the logistic regression curves for the two groups are the same—that is, if $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. If $\beta_{11} = \beta_{12}$ but $\beta_{01} \neq \beta_{02}$, the curves are parallel but not coincident and hence uniform DIF may be inferred. If $\beta_{01} = \beta_{02}$ but $\beta_{11} \neq \beta_{12}$, the curves are not parallel and hence the presence of nonuniform DIF may be inferred.

An alternative but equivalent formulation of the model (1) is

$$P(u = 1) = \frac{e^z}{[1 + e^z]}, \tag{2}$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g). \tag{3}$$

In this formulation, the variable g represents group membership and may be defined as follows:

$$g = \begin{cases} 1 \text{ if examinee is a member of Group 1} \\ 0 \text{ if examinee is a member of Group 2}. \end{cases} \tag{4}$$

The term θg is the product of the two independent variables, g and θ. With the coding given above, the parameter τ_2 corresponds to the group difference in performance on the item, and τ_3 corresponds to the interaction between group and ability. In terms of the parameters of the model in Equation 1,

$$\tau_2 = \beta_{01} - \beta_{02}$$

and

$$\tau_3 = \beta_{11} - \beta_{12}.$$

An item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$ and nonuniform DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$).

In order to see the connection between the logistic regression procedure and the Mantel-Haenszel procedure, we treat the ability variable as discrete with m ability levels (corresponding to a test with $m - 1$ items). We define a variable x_k ($k = 1, \ldots, m - 1$) such that

$$x_k = \begin{cases} 1 \text{ if examinee is a member of ability level k} \\ 0 \text{ otherwise}. \end{cases}$$

All examinees in ability level m receive a score of -1 (or zero). With this coding, z in Equation 2 becomes

$$z = \beta_0 + \sum_{k=1}^{m-1} \beta_k x_k + \tau g. \tag{5}$$

Note that interaction terms between x_k and g are not included. With this formulation, the logistic model can be expressed as

$$\log \frac{P}{(1 - P)} = \beta_0 + \sum_{k=1}^{m-1} \beta_k x_k + \tau g. \tag{6}$$

In this case, $\tau = \log \alpha$ where α is the log-odds ratio defined in Holland and Thayer (1988). Testing the hypothesis that $\tau = 0$ is now equivalent to testing the hypothesis that $\alpha = 1$, the hypothesis tested by the Mantel-Haenszel procedure.

The Mantel-Haenszel procedure can therefore be thought of as being based on a logistic regression model where the ability variable is considered discrete and where no interaction between the ability variable and the group variable is specified. Thus, in the language of experimental design models, the logistic regression model given by Equations 2 and 3 corresponds to an analysis of covariance model while the Mantel-Haenszel model given by Equation 6 corresponds to a randomized block design model. Despite the similarity between the Mantel-Haenszel model and the logistic regression model, it should be noted that the Mantel-Haenszel statistic for testing the hypothesis H_0: $\alpha = 1$ (or $\tau = 0$) was derived using arguments different from those given in the next section.

Estimation and Distribution Theory

Estimation of the parameters $\tau' = [\tau_0\ \tau_1\ \tau_2\ \tau_3]$ for each item is carried out using the method of maximum likelihood. For any item, the likelihood function of the observations is given by

$$L(\text{Data} \mid \tau) = \prod_{i=1}^{N} P(u_i)^{u_i} [1 - P(u_i)]^{1-u_i}. \tag{7}$$

Here, $N = n_1 + n_2$, and

$$P(u_i = 1) = \frac{e^{z_i}}{[1 + e^{z_i}]} \tag{8}$$

where

$$z_i = \tau_0 + \tau_1 \theta_i + \tau_2 g + \tau_3 (\theta_i g). \tag{9}$$

The estimate $\hat{\tau}$ of τ is obtained by maximizing th likelihood function given in (7).

One of the properties of maximum likelihood estimates (MLE) is that they are asymptotically multivariate normal (Bock, 1975), with mean vector τ and variance-covariance matrix Σ, where Σ^{-1} is the information matrix defined as

$$\Sigma^{-1} = -E\left[\frac{\partial^2}{\partial \tau_r \partial \tau_s} \ln L\right] \quad r, s = 0, \ldots, 3. \tag{10}$$

Here E indicates the expectation operator and ln L is the logarithm of the likelihood function given by Equation 7. Thus,

$$\hat{\tau} \sim N(\tau, \Sigma). \tag{11}$$

The asymptotic standard error of the estimate of τ_s (s = 0, . . . , 3) is the square root of the s'th diagonal element of Σ:

$$SE(\hat{\tau}_s) = [\Sigma^{ss}]^{1/2}. \tag{12}$$

Test of Hypothesis

The parameter τ_2 indicates the mean group difference in performance on the item, and τ_3 indicates the interaction between group and ability. If τ_2 is nonzero while τ_3 is zero, we can infer uniform DIF. If τ_3 is nonzero, whether or not τ_2 is zero, we can infer nonuniform DIF. The hypotheses of interest are therefore that $\tau_2 = 0$ and $\tau_3 = 0$. These two hypotheses can be tested simultaneously as

$$H_o: C\tau = \mathbf{0}$$

against

$$H_A: C\tau \neq \mathbf{0}, \tag{13}$$

where

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The statistic for testing the joint hypothesis is

$$\chi^2 = \hat{\tau}'C'(C\Sigma C')^{-1}C\hat{\tau}' \tag{14}$$

which has the χ^2 distribution with 2 degrees of freedom. When the test statistic given by Equation 14 exceeds $\chi^2_{\alpha;2}$, the hypothesis that there is no DIF is rejected. The item can then be flagged for further study by subject matter specialists.

Comparison of the Logistic Regression and Mantel-Haenszel Procedures

As shown earlier, the Mantel-Haenszel procedure can be thought of as being based on a logistic regression model where the ability variable is discrete and no interaction term between the group variable and ability is permitted. In the logistic regression model described in this paper, the ability variable is assumed to be continuous and an interaction term is included. In order to determine if these changes provide an improvement over the Mantel-Haenszel procedure, a comparison study using simulated data was carried out.

The factors manipulated in the study were sample size, test length, and the nature of the DIF; these factors are likely to have the greatest effect on the power of the two procedures. Sample size will affect the power of the logistic regression procedure through its effect on estimation. In small samples, the asymptotic results may not hold and hence the test statistic may not be a valid indicator of the presence of DIF. In the Mantel-Haenszel procedure, small samples may affect the stability of the estimate of the odds ratio in each score group.

Test length affects the accuracy of total score as a measure of ability—the longer the test, the more reliable the total score. Because total score is used as the predictor in the logistic regression model and as the criterion for grouping

examinees in the Mantel-Haenszel procedure, a more reliable score may result in improved estimates of the parameters for both procedures.

The nature (uniform or nonuniform) of the DIF will have an obvious effect on the power of the two procedures to detect the presence of the DIF. In the logistic regression model, the interaction term may adversely affect the power of the procedure when only uniform DIF is present because one degree of freedom is lost unnecessarily. In other words, because the logistic regression procedure is designed to detect nonuniform DIF, it may not be effective in detecting strictly uniform DIF. Conversely, the Mantel-Haenszel procedure is designed to detect uniform DIF and hence may not be effective in detecting nonuniform DIF.

To study the effects of these factors on the relative detection rates of the logistic regression and Mantel-Haenszel procedures, six conditions were simulated: These conditions were obtained by crossing two levels of sample size (250 per group/500 per group) with three levels of test length (40 items/60 items/80 items). Within each test, 20% of the items showed DIF; half were items with uniform DIF, and half were items with nonuniform DIF. (This level of DIF is probably higher than generally found in practice and hence provides a "worst case scenario.") The two DIF detection procedures were compared with respect to the percentage of items with uniform and nonuniform DIF in each condition that was correctly identified and the percentage of items with no DIF that was falsely identified.

Item responses for all items were generated using the program DATAGEN (Hambleton & Rovinelli, 1973) with a three-parameter item response model. In simulating uniform DIF, the discrimination parameters for the two groups were kept the same, and the difficulty parameters were varied to produce the desired degree of DIF. In simulating nonuniform DIF, the difficulty parameters for the two groups were kept the same (and set at a value of 0), and the discrimination parameters were varied. To control the amount of DIF present in each item, the item parameters were chosen such that a prespecified area between the item characteristic curves was obtained. The area between ICCs was calculated using the formula given by Raju (1988). When the item discrimination parameters are the same, as in the case of uniform DIF, and given equal c parameters, this formula reduces to

$$\text{Area} = (1 - c)|b_2 - b_1|.$$

When the item difficulty parameters are the same, as in the case of nonuniform DIF, the formula reduces to

$$\text{Area} = (1 - c)\left|\frac{2(a_2 - a_1)}{Da_1a_2}\ln 2\right|.$$

In simulating uniform DIF, areas of .6 and .8 were chosen; these areas reflect b value differences of .48 and .64 and hence represent moderate to high DIF. In simulating nonuniform DIF, areas of .6 and .8 were also chosen; these areas give nonuniform DIF that is as large as possible within the constraint that the item discrimination values be realistic. Large nonuniform DIF was simulated in order

to give the Mantel-Haenszel procedure an adequate chance to detect the DIF and to determine under what conditions, if any, the Mantel-Haenszel procedure is sensitive to this type of DIF.

In addition to comparing the detection rates of the two procedures under the various conditions, as described above, the power of each DIF detection procedure was studied by carrying out twenty replications of one cell of the design (80 items/500 per group). This cell represented the longest test and largest sample used and thus was the condition under which both procedures should have the best chance of detecting both uniform and nonuniform DIF.

Results

The detection rates for the two procedures are presented in Table 1. For the items with uniform DIF, the two procedures had very similar detection rates, with the Mantel-Haenszel procedure having a very slight advantage. Both were able to detect uniform DIF of this size with about 75% accuracy in samples of 250 per group and with 100% accuracy in samples of 500. For nonuniform DIF, the picture was very different. The Mantel-Haenszel procedure was completely unable to detect nonuniform DIF under any condition. The logistic regression procedure detected nonuniform DIF with about 50% accuracy in small samples and short tests and 75% accuracy in large samples and long tests.

In terms of false positives, the Mantel-Haenszel procedure performed somewhat better than the logistic regression procedure. With a significance level of

Table 1
Comparison of the Logistic Regression and Mantel-Haenszel Procedures in the Detection of Uniform and Non-uniform DIF

Test Length	Type of DIF	No. of Items	Number of Items Flagged as Biased — Sample Size 250 MH	250 LR	500 MH	500 LR
40	Uniform	4	3	3	4	4
	Non-uniform	4	0	2	0	2
	False Positives	32	0	0	0	1
60	Uniform	6	6	5	6	6
	Non-uniform	6	0	3	0	5
	False Positives	48	0	1	0	3
80	Uniform	8	6	6	8	8
	Non-uniform	8	0	4	0	6
	False Positives	64	1	2	0	1

.01, the Mantel-Haenszel procedure consistently produced around 1% false positives under all conditions; the logistic regression procedure produced between 1% and 6% false positives.

Results of the power study are presented in Table 2. In this table, the detection rates are reported as percentages. Because there were 8 items with uniform DIF, 8 with nonuniform DIF, and 20 replications, the percentages were obtained by dividing number of detections by 160 (8 × 20). Similarly, there were 64 items with no DIF; hence, the percentage of false positives was obtained by dividing number of false positives by 1,280 (64 × 20).

Table 2 shows that the two procedures are equally and highly powerful in detecting uniform DIF, but only the logistic regression procedure is able to detect nonuniform DIF with any consistency. Even under the most favorable conditions simulated, the Mantel-Haenszel procedure was unable to detect nonuniform DIF. On the other hand, the false positive detection rate (Type I error rate) for the Mantel-Haenszel procedure was 1%, as expected; the Type I error rate for the logistic regression procedure was 4%, higher than expected.

Conclusion

The logistic regression model described in this paper is more general and flexible than the model that underlies the Mantel-Haenszel procedure. This generality, however, comes at some cost. Computations in the Mantel-Haenszel procedure can be carried out quickly and inexpensively, whereas the logistic regression procedure is iterative and thus more costly. (Our experience has shown that the logistic regression procedure costs 3–4 times as much as the Mantel-Haenszel procedure.)

A comparison between the logistic regression procedure and the Mantel-Haenszel procedure showed that the logistic regression procedure is as powerful as the Mantel-Haenszel procedure in detecting uniform DIF and more powerful

Table 2
Relative Power of the Logistic Regression and Mantel-Haenszel Procedures for Test Length of 80 and Sample Size of 500 Over Twenty Replications

		Percent of Items Flagged as Biased	
Type of DIF	Items x Reps	MH	LR
Uniform	160	96	94
Non-uniform	160	1	71
False Positives	1280	1	4

$\alpha = .01$

than the Mantel-Haenszel procedure in detecting nonuniform DIF. It should be noted that the nonuniform DIF simulated in this study represents disordinal interaction between ability and group membership. This means that the item characteristic curves on which the data were based crossed in the middle of the ability range. It is possible that nonuniform DIF that represents ordinal interaction may occur. In this case, the ICCs will cross at either the low end or the high end of the ability scale, resulting in DIF which may appear to be uniform over most of the range. The Mantel-Haenszel procedure can be expected to be more powerful in detecting this type of nonuniform DIF than the nonuniform DIF simulated here. An investigation of this issue is presently being carried out.

The major advantage of the logistic regression procedure is that it provides a model-based approach for studying DIF. For example, it is possible to include in the model given in Equation 3 curvilinear terms of the ability variable (or transformations of it) as well as other factors that are considered relevant and important. Through this approach, we may be able not only to identify items that are functioning differentially but also to gain a better understanding of the nature of DIF.

References

Bennet, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24,* 41–55.

Bock, R. D. (1975). *Multivariate statistical methods.* New York: McGraw-Hill.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313–334.

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. *Behavioral Science, 17,* 73–74.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston, MA: Kluwer-Nijhoff.

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

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–108.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Shepard, L., Camilli, G., & Averill, M. (1981). A comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317–375.

Spray, J., & Carlson, J. (1986, April). Comparison of loglinear and logistic regression models for detecting changes in proportions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Authors

HARIHARAN SWAMINATHAN is Professor, School of Education, University of Massachusetts, Amherst, MA 01003. *Degrees:* BS, Dalhousie; MS, MEd, PhD, University of Toronto. *Specializations:* statistics, psychometrics, item response theory, and evaluation.

H. JANE ROGERS is Assistant Professor, Teachers College, Columbia University, New York, NY 10027. *Degrees:* BA, MEd, University of New England, Australia; PhD, University of Massachusetts. *Specializations:* measurement, item response theory, statistics, and evaluation.