Using the SAS® System to Detect Differential Item Functioning

Jim Penny, University Research Associates, Jamestown , NC

ABSTRACT

Differential Item Functioning (DIF) is defined by Camilli & Shepard (1994) as the phenomenon where an achievement test item is found to be differentially difficult for the members of a focal demographic group when compared to otherwise identical members of a reference demographic group. For example, some items are found to be more difficult for black males than for equally-abled white males.

There are two general types of DIF: uniform and non-uniform. Uniform DIF is differential functioning that is consistent in direction across varying levels of ability. That is, the item uniformly favors one demographic group over another across all levels of the ability tapped by the item. Non-uniform DIF, the second type, is differential functioning where the direction of the difference changes as ability changes. For example, an item exhibiting non-uniform DIF may be found to favor white males of lower ability, while favoring black males of higher ability.

This paper presents two statistical procedures that are often used to detect DIF and that are easily computed using the FREQ and LOGISTIC procedures of the SAS® System. In addition to a short theoretical review of the procedures, the paper will provide syntax, output, and estimated run-time from analyses of Monte-Carlo data where DIF is admitted by the systematic manipulation of Item Response Theory parameters (Lord, 1980).

The first method, the Mantel-Haenszel procedure (Mantel & Haenszel, 1959), is the most commonly used by major testing agencies. It provides both a χ^2 statistic and a second index known as the "common odds ratio." The odds ratio, which ranges from 0 to positive infinity, is frequently transformed using the natural logarithm to place it on a more interpretable scale centered on 0 that ranges from minus infinity to positive infinity. The SAS FREQ procedure provides an option to compute these indices.

The second method uses logistic regression (Swaminathan & Rogers, 1990), and is easily implemented using the SAS LOGISTIC procedure. One feature of logistic regression when used to detect DIF is that it is sensitive to non-uniform DIF, while the Mantel-Haenszel procedure can be completely insensitive to DIF of this nature; however, logistic regression is more expensive to compute than are the Mantel-Haenszel statistics, and it may be perceived by some practitioners as more difficult to interpret. A second feature of logistic regression, one that contributes to both the computational expense and the initial difficulty of interpretation, is that it is model based, providing parameters to describe items performance and model fit indices to describe the congruence of the data and the model; and as expected, each parameter and index is accompanied with an estimate of statistical significance.

Introduction

There is little in today's society that exceeds the ubiquity and the importance of mental tests, educational measurements, and programmatic assessments in education, in work, and in professional licensure. Students must pass cognitive examinations of many varieties, ranging from teacher-made pop quizzes to nationally normed achievement tests, in order to move from one instructional level to the next; work force promotions, frequently predicated on judgement by professional review, often are based in part upon pencil-and-paper test performance; admission to professional standing, even after extensive academic and practical preparation, is commonly granted only after the candidate's performance on a final cognitive examination has been deemed sufficient by a review board. Even that adolescent right of passage, the driver's license examination, usually includes a pencil-and-paper test of driving knowledge; failure to meet a minimum cut score on this test may deny the young citizen the right to drive.

Given the importance of testing and assessment in the lives of individuals, it is without question that the tests, and the items that comprise these tests, must be reasonably free of bias against the various sub-groups of our population. While the examination of discriminatory test performance has been the subject of much study for the past 3 to 4 decades, the analysis of individual items for evidence of aberrant behavior, or differential item functioning (DIF), between two demographic sub-groups has been the focus of many educational measurement researchers for the past 15 years. Indeed, hardly an issue of any major educational measurement is received without at least one article on differential item functioning, and nearly all of the major educational conferences have entire sections dedicated to DIF and its identification.

Definition of DIF

Differential Item Functioning (DIF) is defined by Camilli & Shepard (1994) as the phenomenon where an achievement test item is found to be differentially difficult for the members of a focal demographic group when compared to otherwise identical members of a reference demographic group. For example, some items are found to be more difficult for black males than for equally abled white males; other items may favor female subjects over male even after the subjects have been equated on ability. It is important to note here that this phenomenon exists even after the items have passed rigorous judgmental review. These are not items that contain offensive language, present stereotypical situations, or make use of specialized knowledge; rather, they are items that for all appearances should represent a common core of knowledge, high school vocabulary items for instance.

It is also important to note at this point that DIF is a statistical definition used to identify items that perform aberrantly from one group of examinees to another; there is no attempt to explain the source of that differential functioning. To that end, studies of DIF are not studies of bias, though many authors use the two terms interchangeably. The term *bias* is used only for items that not only exhibit DIF, but also have been found *invalid* for members of a particular demographic subgroup.

There are two general types of DIF: uniform and non-uniform. Uniform DIF is differential functioning that is consistent in direction across varying levels of ability. That is, the item uniformly favors one demographic group over another across all levels of the ability that is tapped by the item. For instance, an item exhibiting uniform DIF may consistently, or uniformly, favor black subjects over white for all levels of ability. Non-uniform DIF, the second type, is differential functioning where the direction of the difference changes as ability changes. For example, an item exhibiting non-uniform DIF may be found to favor white subjects of lower ability, while favoring black subjects of higher ability.

Methods of DIF Identification

The most commonly cited method used for the detection of DIF is the Mantel-Haenszel procedure (Mantel & Haenszel, 1959). This procedure examines K 2x2 contingency tables where (1) K is the number of possible scores on the test (usually 0 through n, where n is the number of items), (2) the rows are based on demographic group membership, and (3) the columns are based on whether or not the item is answered correctly. The null hypothesis of no association between demographic group membership and probability of correct item response is tested by both a χ^2 statistic and a second index known as the "common odds ratio." The odds ratio is the ratio of odds

for a correct response by members of the two comparison groups at each possible test score. The common odds ratio is the average of the K odds ratios, and is frequently transformed using the natural logarithm to place it on a more interpretable scale.

Neither the Mantel-Haenszel χ^2 statistic nor the common odds ratio are sensitive to non-uniform DIF, and unfortunately, this form of differential functioning arises in studies of DIF with alarming frequency. Some (Green, 1991) have argued in the light of not one single cognitive explanation of non-uniform item response that non-uniform DIF is but a statistical artifact of item analysis and that it is not interpretable from the framework of current educational psychology. However, despite the lack of theoretical underpinning to support its existence, there is no compelling reason to not admit the possibility of non-uniform DIF; indeed, hardly an item analysis exists where non-uniform differential functioning is not found; to that end, measurement specialists must pursue statistical indices of non-uniform DIF.

Another method of detecting DIF, both uniform and non-uniform, uses logistic regression (Swaminathan & Rogers, 1990). Using (1) an ability proxy, (2) group membership, and (3) ability-by-membership interaction, the logistic ogive is fit to the dichotomous item responses. In nearly every case, ability is a statistically significant predictor of item response, and this is exactly as it should be; ability, not group membership or anything else, should affect test and item performance. However, if uniform DIF is present and this differential functioning is associated with group membership, then the group membership variable will most likely be a significant predictor. As well, if group membership is associated with some level of differential functioning that changes as ability changes, then non-uniform DIF is present and the interaction term is likely to be statistically significant.

In comparative studies (Clauser, Nungester, Mazor, & Ripey, 1993; Mazor, 1993; Rogers, 1989; Swaminathan & Rogers, 1990), logistic regression has performed equally well as the Mantel-Haenszel procedure in the presence of uniform DIF. When non-uniform DIF is present, logistic regression performs far better. In the more realistic case where a mixture of uniform and non-uniform DIF is present, one would expect that the relative efficacy of logistic regression over the Mantel-Haenszel procedure as an indicator of differential functioning would depend on the amount of each type of DIF that is present. Items that are uniformly aberrant would be identified equally well by both procedures, barring problems with the model fit in logistic regression. As the amount of non-uniform functioning increased, so would the likelihood that logistic regression would identify the item as aberrant while the Mantel-Haenszel procedure would not.

Unfortunately, logistic regression does have its drawbacks. The first of these is that it is far more computationally expensive than the Mantel-Haenszel procedure. In a recent study involving a Monte-Carlo simulation of 1000 subjects, 500 in each group, taking a 100 item test, approximately 5 hours of CPU time were required on a 33 megahertz 80486DX processor with 8 megabytes of RAM to perform 100 replications of the Mantel-Haenszel procedure using the DOS version of the SAS System. The corresponding analysis using logistic regression required some over 30 hours. Even with a better choice of software, Windows instead of DOS, and a faster processor with more memory, there will still be substantial differences in execution time. However, the time to study 100 test items for 1000 subjects using logistic regression on an 8 megabyte DX2 is only a very few minutes; this does not seem prohibitive, especially given the improved modeling and increased information provided by logistic regression.

A second drawback is that logistic regression may be perceived by some practitioners as more difficult to interpret. This perception is probably more a matter of familiarity with the analysis than of conceptual difficulty. The study of contingency tables occurs early in statistic courses; the study of logistic regression occurs much later in advanced classes, if it occurs at all. However, the principles of interpretation studied in ANOVA and linear regression are easily generalized to logistic regression; the practitioner who is well-founded in traditional statistical procedures should have little or no trouble learning to interpret the results of logistic regression analyses.

DATA Step Requirements

The DATA step requirements to prepare for either analysis, the Mantel-Haenszel procedure or logistic regression, are straight-forward. Variables for the item responses, demographic group membership, and ability must be created and defined for both analyses. If non-uniform DIF is expected, then a variable representing the interaction between demographic group membership and ability must be created for use with logistic regression. All of these variables should be of type INTEGER, though some exceptions are possible.

Item Responses

The dichotomous response vectors of each subject, or observation, for each item should be kept in separate integer variables. For example, if there are 100 items on a test, then one might use ITEM1-ITEM100 to record the item responses. The use of ITEM in the variable name is not required. Another choice might be Q1-Q100, or I1-I100. Of course, there is also no requirement to use enumerated variables; one could just as well choose a separate name for each item, though this choice could result in substantially increased typing. The values for the item responses, regardless of name, are usually coded 0 for incorrect and 1 for correct. Certainly other choices of values are permissible, but this choice permits easy computation of total score by summing across the items.

Group Membership

Demographic group membership should be recorded as an integer for each observation. For the purposes of the Mantel-Haenszel procedure, it does not matter that an integer variable is used; any variable type with discrete values that is acceptable in the TABLES statement of the FREQ procedure will do. Logistic regression is another matter, and will require at least a numerical type; this is especially true if an interaction term between ability and group membership is to be computed.

Variable names for group membership are usually words such as SEX, GENDER, or RACE. Certainly others are possible, and anything that meets the requirements of the SAS System is valid. For the purposes of this paper, group membership will be denoted by the variable name GROUP, with values 0 and 1 where 0 represents the reference group and 1 the focal group. There is no requirement placed on the order or magnitude of these numbers, only that they are discrete with just two possible values.

Ability

Ability is a latent trait that is represented by a proxy measure. This can be a factor score, an independent assessment, or an exotic composite. All that is required is a unidimensionality, but this is critical. If abilities along dimensions other than that which the proxy purports to assess are contributing to the measure, then the efficacy of both the Mantel-Haenszel procedure and logistic regression as indicators of DIF may, and probably will, be compromised. Unfortunately, the assessment of test and item dimensionality is beyond the scope of this paper; however, it is prudent to mention here that linear factor analysis using either oblique or orthogonal rotations has been the preferred method of assessing dimensionality for many years (Lord, 1980), though non-linear methods (Nandakumar, 1994) are gaining respect in the measurement community.

In most cases, ability is represented as the number of correct item responses. If the response vector is held in variables ITEM1-ITEM100, for example, and the responses were coded 0 for incorrect and 1 for correct, then the score on the 100 item test could be computed using SUM(OF ITEM1-ITEM100) and then stored in a variable called SCORE. The choice of variable name for total score is arbitrary, though good coding practice calls for something mnemonic and meaningful.

Interaction

The interaction of ability and group membership, indicative of non-uniform differential functioning, can be assessed only with logistic regression; the Mantel-Haenszel procedure is completely befuddled by it. The

syntax of the LOGISTIC procedure of the SAS System does not allow for the specification of an interaction term on the MODEL statement in the manner of the GLM procedure; rather, the interaction must be computed in a prior data step, and stored in a numerical variable. For this paper, the variable name INTERACT was chosen and then defined as the product of group membership and ability. That is, INTERACT = GROUP*SCORE.

Examples of SAS Syntax

The DATA Step

The precise syntax of the SAS DATA step that defines the item data that are to be studied for DIF depends entirely on the format of the input data set. For this example, I'll assume a fairly simple format of ASCII data with 101 space delimited values per line. The first value indicates the demographic group membership; the last 100 values are the dichotomous item responses. The variable names are GROUP and ITEM1-ITEM100. There is one line per subject. The expected values of group membership and the item responses are 0 or 1, making the length of each line 201 bytes. The proxy for ability is the number of correct items and is kept in the variable SCORE, while the group membership by ability interaction term, to be used only in the LOGISTIC procedure, is computed as the product of SCORE and GROUP and kept in the variable INTERACT. An example of such a data step follows.

```
Data look4dif;
     Infile example;
     Input group item1-item100;
     score = sum(of item1-item100);
     /* INTERACT only for use
        with LOGISTIC */
     interact = score*group;
Run;
```

PROC FREQ

PROC FREQ is used to perform the Mantel-Haenszel analysis. Of the statistics generated, two, perhaps three, are particularly important for those studying DIF. The first is the χ^2 test for general association; the second is the common odds ratio. The third is the Breslow-Day statistic testing the odds ratio for homogeneity across levels of ability. If the odds ratio is found to vary significantly from one level of ability to another, then the validity of its average over ability levels, the common odds ratio, as an index of differential item functioning may be questioned.

By default, PROC FREQ will generate voluminous output for the K 2x2 tables that it examines. To avoid that, the NOPRINT option is used on the TABLES statement. The CMH option on the same statement requests the Cochran-Mantel-Haenszel statistic. An example of this procedure follows.

```
Proc FREQ Data=look4dif;
     Tables score*group*(item1-item100)
      / cmh noprint;
Run;
```

PROC LOGISTIC

PROC LOGISTIC of the SAS System is used to perform the logistic regression. This procedure is sensitive to non-uniform DIF, so the interaction of group membership and ability can be incorporated into the model as appropriate; however, if there is no a priori reason to expect the existance of non-uniform DIF, and this can be the case in many Monte-Carlo studies or follow-up analyses where prior study has shown that non-uniform DIF is not present, then the interaction term can be omitted from the MODEL statement for reasons of parsimony if not analytical propriety. Additionally, Camilli and Shepard (1994) recommend a stepwise approach to logistic regression so that the χ^2s describing the changes in model fit as variables are added to the model can be assessed; they also compare these values to those generated by the Mantel-Haenszel procedure.

The example that follows describes the full model, one that incorporates both uniform and non-uniform DIF; it can be modified to provide a stepwise approach, either forward or backward, as required by the analyst. It should be noted here that a separate analysis must be run from each item; PROC LOGISTIC does not allow the specification of multiple dependent variables as do procedures GLM and FREQ.

```
Proc LOGISTIC Data=look4dif;
     Model item1 = score group interact;
Run;
```

Sample Output with Uniform DIF

Data Source

The data for this example are from a Monte-Carlo study used in other on-going research involving comparisons of logistic regression and the Mantel-Haenszel procedure as indicators of differential functioning that is defined as systematic perturbations of item response theory parameters. For an example of completely uniform DIF, we chose an item that is approximately 3 standard deviation easier for members of the focal group. There are 1000 subjects in each comparison group. This is perhaps too many for use with the Mantel-Haenszel procedure; it is sensitive to sample size, and may be overly susceptible to Type I errors with large samples. However, since this example is for illustrative purposes only, a Type I error is inconsequential.

PROC FREQ

This example is of an item that is easier for members of the focal group. The value of the χ^2 statistic produced by the Mantel-Haenszel procedure is 197.464, with 1 degree of freedom indicating that the item performs statistically significantly differently for members of the focal group. Recall that this is not sufficient evidence for item bias; rather, it is only evidence that the item should be flagged for review by content specialists. See Table 1.

PROC LOGISTIC

This is the same item that was used in the last example. Given that it is known that this item bears only uniform DIF, the interaction term was not included in the analysis. A prior analysis, not shown here, that included only the ability variable in the model statement resulted in a goodness-of-fit index (-2 LOG L) of 819.001. The value for the shown model is 574.579, for a difference of 244.422 with 1 degree of freedom. These statistics are asymtoptically distributed as χ^2; hence, not only is the group membership variable found to be a significant explanatory variable, but its addition to the model results in an overall statistically significant improvement in model fit. See Table 2.

Sample Output with Non-Uniform DIF

Data Source

These data come from the same source as the prior examples involving uniform DIF. The only difference is that the chosen item is extremely poorly discriminating, instead of far easier, for members of the focal group.

PROC FREQ

This is an example of an item that does not discriminate very well for members of the focal group. The χ^2 statistic resulting from the analysis is 0.047 with 1 degree of freedom; it is not statistically significant. This is expected as the item is designed to represent DIF that is purely non-uniform, and the Mantel-Haenszel procedure is completely insensitive to this type of DIF. Additionally, the value of the common odds ratio derived from this analysis is 0.979 and the 95 percent confidence interval includes 1.0, indicating ostensibly nearly equal odds between comparison groups; however, the Breslow-Day statistic provides evidence that the odds ratio probably varies across ability groups, warranting caution in the interpretation of the odds ratio

derived from these data. Again, this is the result of the Mantel-Haenszel procedure being unable to detect non-uniform DIF. See Table 3.

PROC LOGISTIC

This example uses the same item as the last; that is, the item exhibits purely non-uniform DIF. Given the a prior knowledge of the kind of DIF expected with this item, only the ability measure and interaction were included in the model statement; the group variable was omitted. A prior analysis, not shown here, that included only the ability variable resulted in a goodness-of-fit index (-2 LOG L) of 2412.284. The value for the shown model is 2412.284, for a difference of 13.649 with 1 degree of freedom. Again, these statistics are asymtoptically distributed as χ^2; hence, not only is the group membership variable found to be a significant explanatory variable, but its addition to the model results in an overall statistically significant improvement in model fit. See Table 4.

Summary

The Mantel-Haenszel is the most commonly cited index of differential item functioning in use today. It is easily interpreted, easily explained, and economical to compute; however, it is not sensitive to non-uniform DIF, and even without a framework of cognitive psychology to explain its causes, non-uniform DIF is sufficiently common to compel measurement methodologists to examine indicators of differential functioning that are sensitive to both uniform and non-uniform DIF.

Logistic regression is poised at least to supplement, if not supplant, the Mantel-Haenszel procedure as the leading non-parametric indicator of differential item functioning. However, practitioners may not always approach logistic regression with the same degree of comfort with which they approach the study of continency tables; yet with some practical experience, the interpretive skills used in ANOVA and linear regression can be extended to logistic regression, resulting in practitioners who are comfortable with the interpretation of logistic regression results in studies of differential item functioning.

References

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Clauser, B. E., Nungester, R. J., Mazor, K. M., & Ripkey, D. (1993). *Detection of differential item functioning using the Mantel-Haenszel and logistic regression procedures*. In review.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mazor, K. M. (1993). *An investigation of the effects of conditioning on two ability estimates in DIF analyses when the data are two dimensional*, Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses – Comparison of different approaches. *Journal of Educational Measurement, 31*, 17-35.

Rogers, H. J. (1989). *A logistic regression procedure for detecting item bias. Unpublished doctoral dissertation*, University of Massachusetts, Amherst.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 4*, 261-370.

Table 1: Sample output from PROC FREQ when DIF is uniform.

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
3	General Association	1	197.464	0.000

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
(Odds Ratio)	Logit *	0.094	0.060	0.146

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square = 48.378 DF = 55 Prob = 0.724

Table 2: Sample output from PROC LOGISTIC when DIF is uniform.

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	1045.597	580.579	.
SC	1050.505	595.302	.
-2 LOG L	1043.597	574.579	469.018 with 2 DF (p=0.0001)
Score	.	.	424.173 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCEPT	1	-4.0414	0.6989	33.4398	0.0001	.	0.018
SCORE	1	-0.0646	0.00556	134.6207	0.0001	-0.699820	0.937
GROUP	1	3.6850	0.3427	115.6201	0.0001	1.016340	39.846

Table 3: Sample output from PROC FREQ when DIF is non-uniform.

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
3	General Association	1	0.047	0.828

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
(Odds Ratio)	Logit *	0.965	0.778	1.198

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square = 153.586 DF = 80 Prob = 0.000

Table 4: Sample output from PROC LOGISTIC when DIF is non-uniform.

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	2667.836	2418.284	.
SC	2673.437	2435.087	.
-2 LOG L	2665.836	2412.284	253.552 with 2 DF (p=0.0001)
Score	.	.	241.466 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCEPT	1	1.4509	0.1367	112.6862	0.0001	.	4.267
SCORE	1	-0.0239	0.00319	56.2632	0.0001	-0.304286	0.976
INTERACT	1	-0.00596	0.00162	13.4967	0.0002	-0.158601	0.994

ACKNOWLEDGMENTS

The author would like to thank Greensboro College for its support during the preparation and presentation of this paper.

Jim Penny, Ph.D.
University Research Associates, Inc.
116 Tangle Drive
Jamestown, NC 27282

910-454-4416 (Voice and FAX)

Pennyj@iris.uncg.edu