Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 367 – 373
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p367

DIFFERENTIAL ITEM FUNCTIONING CLASSIFICATION FOR POLYTOMOUSLY SCORED ITEMS

Silvia Golia*

Department of Economics and Management, University of Brescia, Italy

Received 30 July 2012; Accepted 06 November 2012
Available online 16 November 2012

Abstract: *Differential item functioning (DIF) refers to a difference in the way a test item functions for comparable groups of test takers. When there are DIF items in the test, it is necessary to make decisions concerning items revision or removal which cannot be based only on the outcome of DIF hypothesis testing statistic; a interpretable measure of the amount of DIF can help in decisions regarding DIF items. Due to the increasing use of polytomous item formats, this paper reviews the systems of DIF classification for polytomously scored items.*

Keywords: *DIF effect size, polytomous items, Liu-Agresti estimate, Signed Area, Standardized Mean Difference, polytomous SIBTEST, ordinal logistic model.*

1. Introduction

Differential item functioning (DIF) refers to a difference in the way a test item functions for comparable groups of test takers. Formally defined, an item displays DIF if subjects of equal proficiency, or equal ability level, on the construct intended to be measured by a test, but from separate subgroups of the population, differ in their expected score on this item ([14]). In DIF analysis, the population is typically divided in two subgroups named reference and focal group; the reference group provides a baseline for performance and the focal group is the focus of fairness concerns. Two types of DIF can be identified and denoted as uniform and nonuniform DIF ([13]). Uniform DIF (UDIF) occurs when the relative advantage of one group over another on a test item is uniform, favoring only one group consistently across the entire scale of ability. Nonuniform DIF (NUDIF) exists when the conditional dependence of group membership and item performance changes in size but not in direction across the entire ability continuum (unidirectional DIF) or when the conditional dependence changes in direction across the entire

* E-mail: golia@eco.unibs.it

ability continuum (crossing DIF). So, NUDIF implies that there is an interaction between ability level and group membership. When DIF items are found in the test, it is necessary to make decisions concerning items revision or removal which cannot be based only on the outcome of DIF hypothesis testing statistic; an interpretable measure of the amount of DIF can help in decisions regarding DIF items. The increasing use of polytomous item formats has led to the development of numerous methods for assessing DIF in polytomous items; therefore, it becomes necessary to identify systems of DIF classification based on DIF effect size appropriate for the polytomuos case to place side by side these methods. This paper wants to review the available classification criteria for the level of DIF in polytomously scored items.

2. Systems of DIF classification

The first system of DIF classification for dichotomously scored items, was developed at Educational Testing Service (ETS), making use of the Mantel Haenszel (MH) delta difference statistic *MH D-DIF* = –2.35 ln[α_{MH}], which is a transformation of the MH constant odds ratio α_{MH} ([9]; [4]; [17]). On the basis of this statistic, an item is classified into three categories: category A contains items with negligible DIF, category B contains items with slight to moderate values of DIF and category C contains items with moderate to large values of DIF. In order to compute α_{MH}, one requires that reference and focal group are matched according to a matching variable X divided in K levels, i.e. an explanatory variable able to form comparable groups such as, for example, the total score. Let P_{Gjk} be the probability of scoring j = 0, 1 for a subject belonging to the group $G = R, F$ (R stands for reference and F for focal group) and level of the matching variable k. The data used in the MH method are in the form of K 2x2 contingency tables, with cell frequencies denoted by N_{Gjk} and total number of individuals at level k denoted by N_k. The estimator of MH constant odds ratio is given by:

$$\hat{\alpha}_{MH} = \sum_{k=1}^{K} N_{R1k} N_{F0k} / N_k \Big/ \sum_{k=1}^{K} N_{F1k} N_{R0k} / N_k = \sum_{k=1}^{K} P_{R1k} P_{F0k} N_k \Big/ \sum_{k=1}^{K} P_{F1k} P_{R0k} N_k .$$

The parameter α_{MH} is assumed constant over all the levels of the matching variable. The ETS classification scheme in terms of $\ln[\hat{\alpha}_{MH}]$is the following: (A) if either $\ln[\hat{\alpha}_{MH}]$is not significantly different from zero or $|\ln[\hat{\alpha}_{MH}]| < 0.43$, (B) if $\ln[\hat{\alpha}_{MH}]$is significantly different from zero and $|\ln[\hat{\alpha}_{MH}]| \geq 0.43$and either: (i) $|\ln[\hat{\alpha}_{MH}]| < 0.64$or (ii) $\ln[\hat{\alpha}_{MH}]$is not significantly greater than 0.43 and (C) if $|\ln[\hat{\alpha}_{MH}]|$is significantly greater than 0.43 and $|\ln[\hat{\alpha}_{MH}]| \geq 0.64$.
Starting from this classification scheme, several authors proposed different tests and statistics to identify the presence and the magnitude of DIF in dichotomously scored items. Given the increasing interest and use of polytomous item formats, this paper wants to review the systems of DIF classification for polytomously scored items. These can be grouped according the approaches described in [13], that is: common odds ratio, item response theory (IRT), mean-difference and logistic regression approach.

2.1 The common odds ratio approach

An approach to measuring DIF effect size is to compute a cumulative common odds ratio. This is the extension to the polytomous case of the analogous approach for dichotomous items which considers the between-group differences in the odds of correct response after conditioning on the observed test score. [12] introduced in the DIF context the Liu-Agresti cumulative common odds ratio α_{LA} ([7]) as a DIF detection method. For a studied item with $J+1$ ordered categories ($j = 0, 1, \ldots, J$) and a matching variable X with K levels, the data for the reference and focal groups at level k ($k = 1, \ldots, K$) of X can be organized in a 2xJ+1 table, with cell frequencies denoted by N_{Gjk} with group marginals denoted by N_{Gk} and category response marginals by N_{jk}. Let the total number of individuals at level k be N_k. Let us denote the cumulative frequency for response level j at stratum k for the reference group $N^*_{Rjk} = N_{R0k} + N_{R1k} + \ldots + N_{Rjk}$ and that for the focal group $N^*_{Fjk} = N_{F0k} + N_{F1k} + \ldots + N_{Fjk}$; the Liu-Agresti estimator of the cumulative common odds ratio is given by:

$$\hat{\psi}_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=0}^{J-1} A_{jk}D_{jk}/N_k}{\sum_{k=1}^{K}\sum_{j=0}^{J-1} B_{jk}C_{jk}/N_k}, \text{ where } A_{jk} = N^*_{Rjk},\ B_{jk} = N_{Rjk} - N^*_{Rjk}, C_{jk} = N^*_{Fjk} \text{ and } D_{jk} = N_{Fjk} - N^*_{Fjk}.$$

Given that $\hat{\psi}_{LA}$ reduces to the invers of $\hat{\alpha}_{MH}$ when $J = 1$, $\hat{\psi}_{LA}$ can be seen as a generalization of $\hat{\alpha}_{MH}$. In order to produce an estimator that is consistent with the scale of $\hat{\alpha}_{MH}$, the estimator $\hat{\psi}_{LA}$ is transformed as $\hat{\alpha}_{LA} = 1/\hat{\psi}_{LA}$. [10] used this $\hat{\alpha}_{LA}$ estimator to propose a categorization scheme for DIF in polytomous items that parallels the ETS scheme for dichotomous items by replacing $\hat{\alpha}_{MH}$ with $\hat{\alpha}_{LA}$ in the ETS scheme. The classification scheme in terms of $\hat{\alpha}_{LA}$ becomes: (A) if either $\ln[\hat{\alpha}_{LA}]$ is not significantly different from zero or $|\ln[\hat{\alpha}_{LA}]| < 0.43$; (B) if $\ln[\hat{\alpha}_{LA}]$ is significantly different from zero and $|\ln[\hat{\alpha}_{LA}]| \geq 0.43$ and either: (i) $|\ln[\hat{\alpha}_{LA}]| < 0.64$ or (ii) $|\ln[\hat{\alpha}_{LA}]|$ is not significantly greater than 0.43 and (C) if $|\ln[\hat{\alpha}_{LA}]|$ is significantly greater than 0.43 and $|\ln[\hat{\alpha}_{LA}]| \geq 0.64$.

2.2 The item response theory approach

In the IRT approach, one of the methods for detecting DIF is based on the signed area (SA), that is the area between the expected score functions specified for the reference and focal groups; it is suitable to detect UDIF and unidirectional DIF. For dichotomously scored items, the expected score function is equal to the item response function. Let $E_G(Y|\theta)$ be the expected value of the item response variable Y conditional on the ability level θ for group G; SA is given by $SA = \int_{-\infty}^{+\infty} E_R(Y|\theta) - E_F(Y|\theta)\, d\theta$. For dichotomously scored items, [15] demonstrated that, for the three-parameter logistic model, under the assumption that the guessing parameter c is the same for the two groups, $SA = (1 - c)\ \delta_F - \delta_R$, where δ_G is the difficulty parameter of the item for group G. In the case of the Rasch model (RM), SA becomes $SA = \delta_F - \delta_R$. In this case is possible to use the ETS classification scheme that involves the MH constant log odds ratio $\ln[\alpha_{MH}]$. Let

one consider the α_{MH} based on a single level of the matching variable, say k, this parameter becomes:

$$\alpha_k = P_{R1k}P_{F0k}N_k / P_{F1k}P_{R0k}N_k = \frac{P_{R1k}}{P_{R0k}} \Big/ \frac{P_{F1k}}{P_{F0k}}, \text{ and its logarithm } \ln[\alpha_k] = \ln\left[\frac{P_{R1k}}{P_{R0k}}\right] - \ln\left[\frac{P_{F1k}}{P_{F0k}}\right].$$

Under the hypothesis that the RM is the data generating model, $\ln[\alpha_k] = \delta_F - \delta_R$; the RM and MH approaches are both based on the relative odds of success of the two groups on the suspect item. Starting from this remark, [6] applied the ETS system of DIF classification to the difference of the item difficulty parameters for the two groups. Given that under RM this difference coincides with *SA*, the ETS classification scheme can be applied to *SA* at least under the RM. In the case of polytomously scored items, if the model considered is the Graded Response Model (GRM) or the Partial Credit Model (PCM), [2] and [11] demonstrated that $SA = \sum_{j=1}^{J}(\delta_{Fj} - \delta_{Rj})$. For PCM, the difficulty parameter of each item can be decomposed as $\delta_{Gj} = \delta_G + \tau_j$, where τ_j is the threshold parameter; if DIF affects only the δ_G parameter and leaves unchanged the thresholds, as in the UDIF case, the SA becomes: $SA = J(\delta_F - \delta_R)$, where J is the number of the thresholds. This formula is a scale transformation of *SA* in the dichotomous case, so the system of DIF classification becomes: (A) if $|SA/J| < 0.43$; (B) if $0.43 \leq |SA/J| < 0.64$; (C) if $|SA/J| \geq 0.64$.
This scheme, derived from what found in the dichotomous case for the RM, can be also applied in the case in which the reference model is the GRM, given that it shares the same *SA* formula with the PCM.

2.3 The mean-difference approach

The mean-difference approach evaluates the conditional between-group difference in the expected value of the item response variable *Y*. Two statistics belong to this approach, that is the Standardized Mean Difference (*SMD*) [20] and the polytomous SIBTEST [1].
Given a matching variable with *K* levels, the *SMD* it is defined as:

$$SMD = \sum_{k=1}^{K} w_{Fk}m_{Fk} - \sum_{k=1}^{K} w_{Fk}m_{Rk},$$

where m_{Gk} is the mean item score for the group *G* in the *k*th level of the matching variable, and w_{Fk} is the proportion of focal group members who are at *k*th level of the matching variable. The ETS classification scheme that takes into account *SMD*, involves also the Mantel's chi-square statistic ([8]) and the standard deviation of the item scores for the total group (*SD*). Therefore, based on the following rules (R. Zwick, personal communication, July 20, 2012), an item is classified as: (A) if either Mantel's chi-square is not significantly different from zero or $|SMD/SD| \leq 0.17$; (B) if Mantel's chi-square is significantly different from zero and $0.17 < |SMD/SD| \leq 0.25$ or (C) if Mantel's chi-square is significantly different from zero and $|SMD/SD| > 0.25$.
The polytomous SIBTEST introduced in [1], represents an extension of the SIBTEST ([16]) able to handle polytomous items. It is based on the regression of item performance onto an estimate of ability based on classical test theory of matching variable true score. The differences in the estimated item-true score regressions for focal and reference groups are averaged across same

observed matching test scores with a focal group weighting function. A subset of n items X_i must be presumed approximately unidimensional and DIF-free, in order to estimate the matching test true score $X = \sum_{i=1}^{n} X_i$. For each score level k of X, $k = 0, 1, \ldots, n_H$, the difference of the estimated item-true score regressions for focal and reference groups is $d_k = \overline{Y}_{Rk} - \overline{Y}_{Fk}$, where $\overline{Y}_{Gk}$ is the mean item scores for the G group with score k. The polytomous SIBTEST statistic is given by $\hat{\beta} = \sum_{k=0}^{n_H} p_k d_k$, where p_k is a weight equal to the proportion of all examinees with score k or the proportion of focal group members with score k. Given that the standardization statistic STD P-DIF developed by [5] uses the same DIF scale as the SIBTEST ([16]), it is possible to adopt for SIBTEST the system of DIF classification reported in [4]. Given that the SIBTEST ranges between -1 and 1, before applying this same classification scheme to the polytomous SIBTEST, it is necessary to place $\hat{\beta}$ on the same scale by dividing it by the difference between the maximum and the minimum possible score on the item. The system of DIF classification for polytomous SIBTEST becomes: (A) if $\hat{\beta}$ is not significantly different from zero or $|\hat{\beta}/J| \leq 0.05$; (B) if $\hat{\beta}$ is significantly different from zero and $0.05 < |\hat{\beta}/J| \leq 0.10$ and (C) if $\hat{\beta}$ is significantly different from zero and $|\hat{\beta}/J| > 0.10$.

2.4 The logistic regression approach

The logistic regression approach for dichotomous items can be adapted to the polytomous case using one of the available ordinal logistic models. [18] proposed the use of a proportional odds model; this method allows one to get a measure of DIF making use of an R^2 measure of effect size. Three equations are used in Zumbo's method: a full model (Model III), where the item score is predicted from total score, group membership and the interaction of group membership and total score, an equation modeling UDIF only (Model II), that is the full model minus the interaction term, and a null DIF equation (Model I), where the item score is predicted from only the total score. The generalized coefficient of determination ([3]) is one of the possible R^2 measures. A guideline for interpreting the difference between two R^2 measures R_d^2 is the following [19]: (A) if $R_d^2 < 0.13$; (B) if $0.13 \leq R_d^2 \leq 0.26$; (C) if $R_d^2 > 0.26$. The comparison between Model I and Model III $R_d^2 = R^2(\text{Model III}) - R^2(\text{Model I})$can be seen as a measure of the combined UDIF and NUDIF magnitude, the comparison between Model I and Model II $R_d^2 = R^2(\text{Model II}) - R^2(\text{Model I})$ can be seen as a measure of the magnitude of UDIF, whereas the comparison between these two R_d^2 assesses the magnitude of NUDIF.

3. Conclusions

When DIF items are found in the test, it is necessary to make decisions concerning items revision or removal which cannot be based only on the outcome of DIF hypothesis testing statistic, therefore, a interpretable measure of the amount of DIF can help in decisions regarding DIF items. This paper reviews the systems of DIF classification for polytomous items available in literature. Between them, only the system based on ordinal logistic model is able to give a

measure of the crossing DIF size. Although this type of NUDIF is less frequent than UDIF and unidirectional DIF, many researchers have documented its presence in real test data analysis. Further research is needed on classification schemes for the size of this kind of NUDIF.

Acknowledgement

The author wish to thank L.A. Roussos and R. Zwick for their availability and help and two anonymous reviewers for their useful comments.

References

[1]. Chang, H., Mazzeo, J., Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement,* 33, 333–353.

[2]. Cohen, A.S., Kim, S.-H., Baker, F.B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, 17, 335–350.

[3]. Cox, D.R., Snell, E.J. (1989). *The analysis of binary data* (2nd ed.). New York: Chapman and Hall.

[4]. Dorans, N.J., Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization, in *Differential item functioning*, eds. P.W. Holland and H. Wainer, Hillsdale, NJ: Erlbaum, 35–66.

[5]. 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.

[6]. Linacre, J.M., Wright, B.D. (1989) Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions*, 3(2), 51–53.

[7]. Liu, I-M, Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, 52, 1223–1234.

[8]. 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.

[9]. 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.

[10]. Penfield, R.D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education*, 20(3), 335–355.

[11]. Penfield, R.D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement*, 47, 129–149.

[12]. Penfield, R.D., Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative Common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, 40(4), 353–370.

[13]. Penfield, R.D., Camilli, G. (2007). Differential item functioning and item bias, in *Handbook of statistics: Volume 26. Psychometrics,* eds. S. Sinharay and C.R. Rao, New York, NY: Elsevier, 125–167.

[14]. Roussos, L.A., Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses, in *The Sage handbook of quantitative methodology for the social sciences*, eds. D. Kaplan, Thousand Oaks, CA: Sage, 107–116.
[15]. Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
[16]. Shealy, R., Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159–194.
[17]. Zieky, M. (1993). Practical questions in the use of DIF statistics in item development, in *Differential item functioning*, eds. P.W. Holland and H. Wainer, Hillsdale, NJ: Erlbaum, 337–364.
[18]. Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.
[19]. Zumbo, B.D., Thomas, D.R. (1996). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia.
[20]. Zwick, R., Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, 21, 187–201.