Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation
EJASA:DSS (2010). Vol 1, Issue 1, 16 – 25
e-ISSN 2037-3627, DOI 10.1285/i2037-3627v1n1p16
© 2010 Università del Salento, SIBA http://siba-ese.unile.it/index.php/ejasa_dss

THE ASSESSMENT OF DIF ON RASCH MEASURES WITH AN APPLICATION TO JOB SATISFACTION

Silvia Golia[*]

Department of Quantitative Methods, University of Brescia, Italy.

Received 31 January 2010. Accepted 1 June 2010
Available online 17 December 2010

Abstract: *The present study addresses the issue in assessment of the impact of differential item functioning on the measures obtained applying the Rasch model when the questionnaire is formed by polytomous items. An item is said to display differential item functioning when it behaves differently among different groups of respondents (for example males and females). A simulation study is used in order to deal with the issue. A differential item functioning analysis is performed making use of a real database concerning the Survey on Italian Social Cooperatives carried out in 2007.*

Keywords: *Rating Scale model, uniform differential item functioning, simulation study, ICSI^{2007}*

1. Introduction

When a validated test is used to measure a latent trait, it is important to ensure that the test itself and the items structure are invariant over population characteristics such as, for example, gender or age. Differential Item Functioning (DIF) analysis examines the relationships among item responses, levels of the trait being measured (ability) and subgroup membership. For a given level of trait, the probability of endorsing a specified item response should be independent of subgroup membership; if it does not happen, then that item is said to exhibit DIF. If the questionnaire is made of dichotomously scored items, then an item is said to display DIF if the probability of positive response varies according to group membership. When a test is made of polytomous items, the definition of DIF is more complex. Different patterns of DIF can appear in the data, namely the constant DIF, that is DIF is constant across response categories, unbalanced

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

DIF, where, for example, DIF affects only one category (the lower or the higher score category) and balanced DIF, that is DIF is balanced across score categories [18].

In a typical DIF study, subgroups are studied in pairs, with one group labeled the *reference group* (often the majority) and the other the *focal group*. The term focal refers to the particular group of primary interest (for example ethnic minorities) whereas reference refers to the group to which the focal group item responses are to be compared.

Two types of DIF can be identified: *uniform* and *nonuniform* [8]. *Uniform* DIF (UDIF) occurs when an item is endorsed at a consistently higher level by one group over the other group at all levels of the underlying trait. This situation is referred to as UDIF because the DIF effect is uniform across the latent trait continuum. *Nonuniform* DIF (NUDIF) means that at certain levels of the underlying trait, one group has higher scores, while at other levels the opposite is the case.

Several methods based, for example, on the Mantel-Haenzel procedure [5] or the logistic regression [15] were proposed to asses DIF in dichotomous items. With the increasing use of tests with polytomous items, extensions to the polytomous case of the previous methods ([9], [17], [18]) and new approaches, such as the logistic discriminant function analysis [11], were developed. A review of polytomous DIF methods is given by [13] and [12].

The present simulation study addresses the issue in assessment of the impact of the ignored effect of the presence of UDIF on the measures obtained applying the Rasch model when the questionnaire is formed by polytomous scored items. The findings are then applied to a real database, composed of the responses to the job satisfaction items included in the Survey on Italian Social Cooperatives carried out in 2007 (ICSI^{2007}).

The paper is organized as follows. Section 2 reviews the Rasch model while Section 3 reports the results of the simulation study. Section 4 investigates if the items related to the job satisfaction section of the national survey display UDIF with reference to three variables: gender, cooperative type and membership. Conclusions follow in Section 5.

2. The Rasch model

The *Rasch model* [14] is a model which converts raw scores into linear and reproducible measurement. Its underlying hypotheses are *unidimensionality*, which means that all considered items in the questionnaire measure only a single construct, i.e. the latent trait under study, and *local independence*, which implies that conditional to the latent trait, the response to a given item is independent from the responses to the other items describing the latent trait. The mathematical form of the Rasch model provides the separation of item and person parameters with the consequence that the total score for the items or persons is sufficient statistic for the item or person parameters.

If the data fit the model, then the measures produced applying the Rasch model to the sample data are objective and expressed in logit[1] [16]. The property of specific objectivity means that the relative location of pairs of persons and pairs of items on the underlying continuum are sample independent.

The model formulated by Rasch dealt with dichotomous data; in a paper appeared in 1978, Andrich proposed a model useful for analyzing rating scale data called *Rating Scale Model*

[1] The logit scale is an interval scale.

(RSM) [1]. The model states that the log-odds ratio of two adjacent categories equals to the difference between the person's ability, item difficulty and step calibration. The RSM is expressed as:

$$\ln\left(\frac{P_{nij}}{P_{ni(j-1)}}\right)=\beta_n-\delta_i-\tau_j \qquad j=1,...,m \tag{1}$$

where P_{nij} denotes, for respondent *n*, the probability of scoring *j* on item *i*, whereas $P_{ni(j-1)}$ the probability of scoring *j*-1 on the same item, β_n identifies the *ability* of person *n*, δ_i the *mean difficulty* of item *i* and τ_j, called *threshold*, is the point of equal probability of categories *j*-1 and *j* ($\tau_0 \equiv 0$ and $\sum_{j=1}^{m}\tau_j=0$).

The present study aims at investigating the impact of UDIF on the measures obtained when RSM is used. The case of DIF constant across all the response categories is considered.

In order to simulate UDIF, for the item *i* (1) becomes:

$$\ln\left(\frac{P_{nij}}{P_{ni(j-1)}}\right)=\beta_n-\delta_i-\tau_j-d_i\cdot group_n \tag{2}$$

where the variable $group_n$ is a dummy variable coded as 1 if the subject *n* belongs to the focal group and 0 otherwise. The parameter d_i refers to the difference, for the item *i*, between the reference and focal group mean difficulty parameters, i.e. $d_i=\delta_{ir}-\delta_{if}$, and represents a measure of the DIF size.

3. The Simulation Study

The present Section reports the results of a simulation study which addresses the issue in assessment of the impact of UDIF on the measures obtained applying the RSM.

The data are generated as follows. A sample of 1000 abilities was drawn from a standard normal distribution and attributed at random to the reference and focal groups. These abilities represent the target or *true abilities* β_n and are used to generate the responses to each of the 15 items forming the questionnaire according to (1), when the item is DIF free, and (2) when the item has UDIF.

A set of the 15 difficulty parameters δ_i is drawn from a continuous uniform distribution on the interval from -1.9 to 1.9 and transformed so that the parameters sum is equal to zero, as required by the calibration procedure. The first 12 common items (without any DIF) have mean difficulties δ_i [-1.7684, -1.4726, -0.8373, -0.5323, -0.3092, -0.2662, -0.1237, 0.7783, 0.9029, 1.0733, 1.3682, 1.8274]; Table 1 reports the mean difficulties of the remaining 3 items for the reference group and the values of the parameter d_i, difference between the item mean difficulty parameters for reference and focal group, when slight to moderate (Case A) and moderate to large (Case B) UDIF effects are generated.

Table 1. Mean difficulties of items 13 to 15 for reference group and values of d_i

	Ref. Group	d_i	
Item	δ_i	***Case A***	***Case B***
13	-0.9496	0.34	0.6303
14	0	-0.64	-1.1533
15	0.3092	0.30	0.5230

The set of threshold parameters τ_j is [-1, -0.5, 0, 0.5, 1], which implies six response categories. Two sample size proportions are used; in the first case, the reference group sample size n_r is 492 whereas the focal group sample size n_f is 508 (approximately 50-50), in the second case n_r = 754 and n_f = 246 (approximately 75-25).

For all the four combinations (sample size proportion × UDIF effect), 100 data sets were simulated and analyzed and 100 sets of estimated abilities and item difficulties were computed. In the calibration procedure the response probability is derived from (1) and the analysis was performed by setting the mean of item difficulty estimates to 0.0 logits and by using the (unconditional) maximum likelihood estimation method[2].

All the four situations produce estimates of ability which are not significantly influenced by the presence of the three items with UDIF.

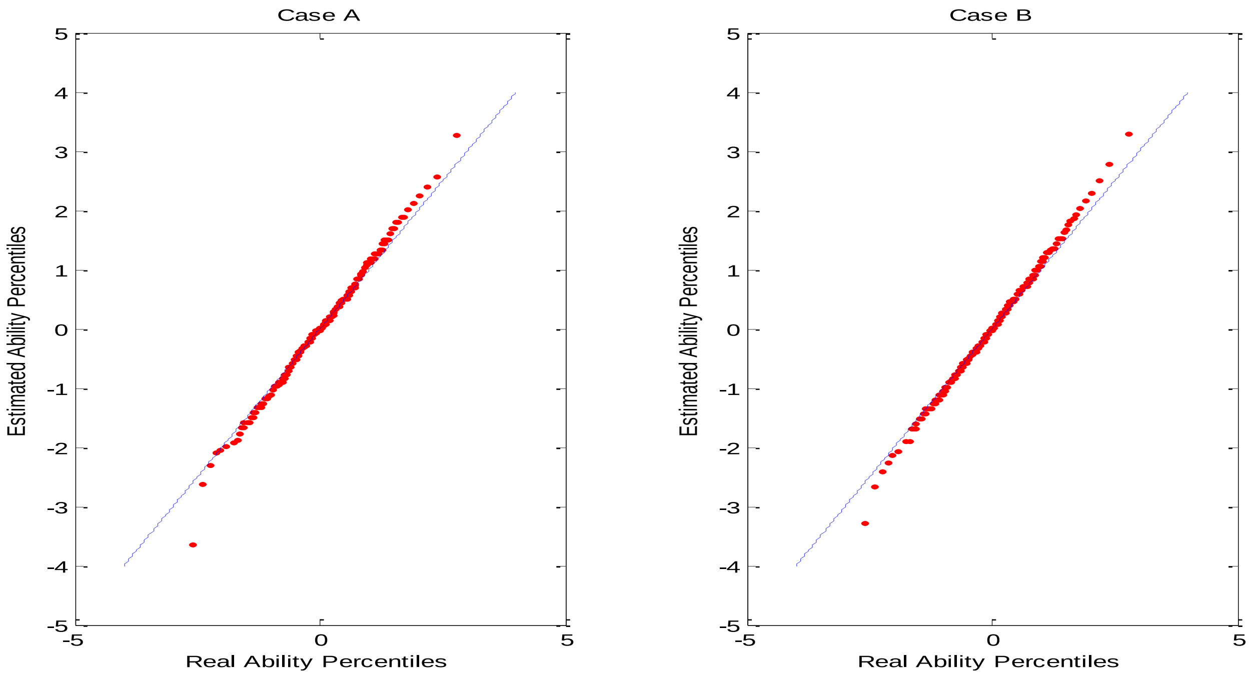

Figure 1. Q-Q plots of true versus estimated abilities. Case A (n_r = 492, n_f = 508) and Case B (n_r = 754, n_f = 246)

[2] The data simulation was performed using Matlab 6.5 whereas the Rasch analysis using Winsteps 3.65.

Concerning true and estimated abilities, making use of Quantile-Quantile plots, as in Figure 1, as well as the two-sample Kolmogorov-Smirnov test (K-S), one can conclude that in almost all the cases the distributions are the same. Similar result appears when the mean absolute bias between true and estimated abilities, calculated on the least able (level of estimated ability lower than the first decile), the most able (level of estimated ability higher than the ninth decile) and the mid-able (level of estimated ability bounded by the first and ninth deciles) subjects, is taken into account, as shown in Table 2. When the items with UDIF are involved in the estimation procedure, the bias is comparable with that observed when all the 15 items are DIF free[3]. Moreover the mean correlation between true and estimated abilities is extremely high.

Table 2. Mean correlation and mean absolute bias between true and estimated abilities (std. in parenthesis)

		Bias		
	Correlation	**Least able**	**Most able**	**Mid-Able**
DIF free	0.963 (0.002)	0.310 (0.056)	0.306 (0.055)	0.222 (0.018)
Case A				
n_r = 492; n_f = 508	0.963 (0.002)	0.287 (0.052)	0.294 (0.050)	0.218 (0.021)
n_r = 754; n_f = 246	0.963 (0.002)	0.299 (0.054)	0.293 (0.055)	0.217 (0.021)
Case B				
n_r = 492; n_f = 508	0.963 (0.002)	0.273 (0.049)	0.271 (0.052)	0.208 (0.018)
n_r = 754; n_f = 246	0.963 (0.002)	0.281 (0.051)	0.279 (0.053)	0.212 (0.019)

When Case A is taken into account, no sign of multidimensionality is found in the data; the mean value of the first eigenvalue of Principal Component Analysis (PCA) on Rasch residuals is consistent with what found in [2], that is 1.278 (std. 0.042). Moreover, the infit and outfit statistics computed for the three items with DIF do not indicate any problematic behaviour. When the DIF size becomes more prominent, as in Case B, a light sign of multidimensionality is found in the data when the respondents are divided in two almost equal groups; the mean value of the first eigenvalue of PCA on Rasch residuals is 1.39 (std. 0.035). The item involved in this light second dimension is item 14, the one with the biggest UDIF size, which is misfitting; mean infit value = 1.346 (std. 0.054) and mean outfit value = 1.347 (std. 0.061). When the reference group is larger than the focal group, this result weakens.

On the basis of this analysis one can conclude that the presence of a low number of items exhibiting UDIF, as in the present study, does not compromise the goodness of the ability estimates.

The following analysis aims at examine what happens if the three items displaying DIF are deleted, reducing the length of the questionnaire. The rejection percentages of the null hypothesis of equal distribution of true and estimated abilities, making use of K-S test, vary between 85% and 90%. The study of the Quantile-Quantile plots of the quantiles of the true abilities versus the quantiles of the estimated ones shows that the extreme subjects, that is the least and most able respondents, are more difficult to estimate, due to the reduced number of items. An example is displayed in Figure 2.

[3] The item difficulties of the reference group are used to simulate the responses of 1000 subjects which form the bases of the results labeled as *DIF free* in Table 2.

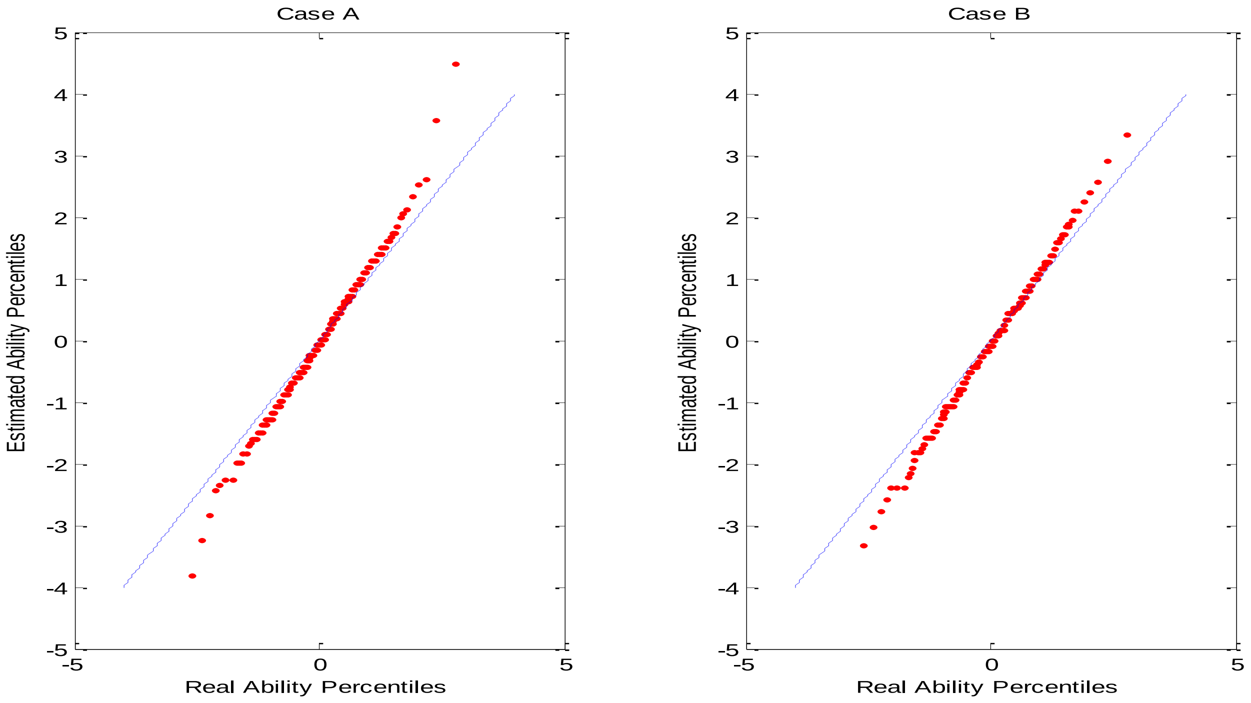

Figure 2. Q-Q plots of true versus estimated abilities when the DIF items were deleted. Case A (n_r = 492, n_f = 508) and Case B (n_r = 754, n_f = 246)

Though the correlation between true and estimated abilities is high, the mean absolute bias computed on the three groups of respondents shown in Table 3 is higher than the one reported in Table 2.

Table 3. Mean correlation and mean absolute bias between true and estimated abilities when the DIF items were deleted (std. in parenthesis)

	Correlation	Bias Least able	 Most able	 Mid-Able
Case A				
n_r = 492; n_f = 508	0.953 (0.003)	0.383 (0.064)	0.328 (0.054)	0.266 (0.029)
n_r = 754; n_f = 246	0.953 (0.003)	0.391 (0.066)	0.324 (0.060)	0.263 (0.028)
Case B				
n_r = 492; n_f = 508	0.954 (0.002)	0.390 (0.068)	0.325 (0.063)	0.262 (0.028)
n_r = 754; n_f = 246	0.953 (0.002)	0.387 (0.069)	0.325 (0.062)	0.264 (0.028)

One can conclude that the elimination of the items with UDIF has a negative effect on the ability estimates, therefore it is convenient to preserve the items in the estimation procedure and use the information coming from a UDIF analysis to describe in a better way the phenomenon under study.

4. Real Data

The present section reports a DIF analysis performed making use of a real data set composed of the responses to the job satisfaction items included in the ICSI^{2007}, which involved 320 Italian social cooperatives of type A and B and 4,134 paid workers [3].

The analysis showed in [4] identified a final job satisfaction scale composed by the 11 items reported in Table 4.

Table 4. The items which compose the job satisfaction scale

Item		How satisfied are you with…
1	*Involv*	your involvement in the Cooperative decisions?
2	*Transp*	the transparency in your relation with the Cooperative?
3	*Coop-Recog*	the recognition by the cooperative of your work?
4	*Growth*	your vocational training and professional growth?
5	*Indep*	your decisional and operative independence?
6	*Career*	your achieved and prospective career promotions?
7	*Fulfil*	your personal fulfilment?
8	*Team*	the relations within the team?
9	*Super*	the relations with your superiors?
10	*Variety*	the variety and creativity of your work?
11	*Coll-Recog*	the recognition by co-workers of your work?

The DIF variables considered are gender (1,068 males and 3,066 females), cooperative type (3,234 workers employed in type A cooperatives and 900 in type B cooperatives) and membership (3,056 members and 989 no members).

In order to detect which item shows DIF, with reference to the three DIF variable, the Mantel test [9] is used. Table 5 reports the list of the items showing DIF with respect to gender, cooperative type and membership variables, difficulty parameters estimate for reference and focal group and two measures of DIF effect size (absolute value), which allow to evaluate the DIF severity.

The cooperative type DIF analysis, as well as the membership DIF analysis highlighted a large number of items displaying DIF. The simulation study reported in [6] has shown that an high number of UDIF items does not compromise the goodness of the estimated abilities.

A descriptive measure of DIF effect size for polytomous items is based on the standardized mean difference (SMD) [19]. The subjects are divided in strata which are formed using a stratification variable, such as the raw test score. The SMD is given by the difference between the mean item score of the focal group and the weighted mean item score of the reference group, where the weights are the proportion of focal examinees in each stratum. In order to obtain a SMD effect size estimate, SMD is divided by the within-group standard deviation of the studied item, pooled over the two groups.

The Educational Testing Service (ETS) developed classification guidelines for the SMD effect size that can be used to determine the practical significance of DIF ([10]; [20])[4].

In order to evaluate the SMD effect size, in the analysis the workers were classified into ten groups using deciles of their raw scores. The DIF contrast in Table 5 is a Rasch equivalent of SMD effect size which uses logit measures and allows for missing data. The commonly-accepted criteria to categorize the severity of DIF (as proposed by ETS) are shown in [7].

Table 5. Difficulty parameters for reference and focal group and DIF effect size estimates

Item	**Reference**	**Focal**	**DIF contrast**	**SMD effect size**
Gender				
	Female	*Male*		
Career	1.19	0.97	0.23	0.12
Super	-0.87	-0.65	0.23	0.08
Team	-0.92	-0.76	0.16	0.06
Cooperative Type				
	Type A	*Type B*		
Variety	-0.41	0.05	0.46	0.22
Career	1.20	0.92	0.28	0.20
Coop-Recog	0.20	-0.04	0.24	0.15
Involv	0.73	0.55	0.17	0.13
Coll-Recog	-0.29	-0.03	0.27	0.12
Transp	0.09	-0.06	0.15	0.10
Membership				
	Member	*No Member*		
Involv	0.64	0.83	0.19	0.10
Coop-Recog	0.19	0.02	0.17	0.10
Super	-0.78	-0.92	0.14	0.07
Transp	0.09	-0.03	0.12	0.07
Career	1.11	1.21	0.09	0.06

With regards to the gender variable, male workers constitute the focal group and the items *Career*, *Team* and *Superiors* display DIF. The difficulty parameters estimates for the two groups allow to observe that male workers find career promotion less difficult to satisfy whereas relations with superiors and within the team more difficult to satisfy than female workers. Nevertheless, the evaluation of the two DIF effect size measures reveals that the three items have negligible DIF.

Workers of type B cooperatives form the focal group when cooperative type variable is taken into account in the DIF analysis. *Variety* is the item with highest DIF size followed by *Carrer* and *Coll-Recog.* The difficulty parameters estimates for the two groups allow to observe that

[4] The ETS system classifies a polytomous item with Mantel's chi-square significant as having negligible DIF if the absolute value of the SMD effect size is less than or equal to 0.17, moderate DIF if the absolute value of the SMD effect size is over 0.17 and less than or equal to 0.25 and large DIF if the absolute value of the effect size is over 0.25.

workers of type B cooperative find career promotion and relations with the cooperative (*Involv, Trasp* and *Coop-Recog*) less difficult to satisfy whereas variety in work and recognition by co-workers of work done more difficult to satisfy than workers of type A cooperative. The measures of the SMD effect size show that only *Carrer* and *Variety* have moderate DIF (the DIF contrast indicates only *Variety*); the other items exhibit negligible DIF.

Involv, Trasp, Coop-Recog, Carrer and *Super* are the items displaying DIF when membership is taken into account (no members constitute the focal group). Nevertheless for all these items the two DIF effect size measures show negligible DIF. From the difficulty parameters estimates one can concludes that no members find recognition by the cooperative of the work done, transparency in the relation with the cooperative and relations with your superiors less difficult to satisfy whereas involvement in the cooperative decisions and career promotion more difficult to satisfy than members.

5. Conclusions

The present study addressed the issue in assessment of the impact of UDIF on the measures obtained applying the Rasch model when the questionnaire is formed by polytomous items. The simulation study considers three items with DIF of different magnitude, a sample of 1000 respondents divided in two groups with equal size, in one case, and different sizes in the second case. On the basis of the obtained results one can conclude that the presence of a low number of items exhibiting DIF does not compromise the goodness of the ability estimates, whereas deleting the items with DIF has a negative effect on the estimates. It is convenient to preserve the items in the estimation procedure and use the information coming from a DIF analysis to describe in a better way the phenomenon under study. The analysis of the real data set composed of the responses to the job satisfaction items included in the ICSI^{2007} has found few items with DIF with reference to the three DIF variables taken into account: gender, cooperative type and membership.

Acknowledgement

The author wishes to thank the two anonymous reviewers; their comments resulted in a significant improvement of the manuscript.

References

[1]. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

[2]. Brentari, E., Golia, S. (2007). Unidimensionality in the Rasch model: how to detect and interpret. *Statistica*, 67(3), 253–261.

[3]. Carpita, M. Eds. (2009). *La qualità del lavoro nelle cooperative sociali. Misure e modelli statistici*, Milano: Franco Angeli.

[4]. Carpita, M., Golia, S. (2008). Subjective measures of quality of work in the Italian social cooperatives. *Rapporti di ricerca del Dipartimento Metodi Quantitativi, Università di Brescia, Facoltà di Economia, 312*.

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

[6]. Golia, S. (2009). The impact of uniform and nonuniform differencial item functioning on Rasch measure. In: *Classification and Data Analysis 2009 Book of short papers*, 517-520. Padova: Cleup.

[7]. Linacre, J.M. (2006). *Winsteps Rasch measurement computer program*. Chicago: Winsteps.com.

[8]. Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105–107.

[9]. Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of American Statistical Association*, 58, 690–700.

[10]. Meyer, J.P., Huynh, H., Seaman, M.A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, 41, 331–344.

[11]. Miller, T.R., Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30, 107–122.

[12]. Penfield, R.D., Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19, 5–15.

[13]. Potenza, M.T., Dorans, N.J. (1995). Evaluation DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23–37.

[14]. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Copenhagen: The Danish Institute of Educational Research.

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

[16]. Wright, B.D., Masters, G.N. (1982). *Rating Scale Analysis*, Chicago: MESA Press.

[17]. Zumbo, B.D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores, *Directorate of Human Resources Research and Evaluation, Department of National Defense, Ottawa.*

[18]. Zwick, R., Donoghue, J.R., Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.

[19]. 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.

[20]. Zwick, R., Thayer, D. T., Mazzeo, J. (1997). Describing and categorizing DIF in polytomous items. *ETS Research Report 97-05, Princeton, NJ.*