Measurement 50 (2014) 186–197

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

A scale purification procedure for evaluation of differential item functioning

Muhammad Naveed Khalid [a,*], Cees A.W. Glas [b]

[a] Cambridge English Language Assessment, University of Cambridge, 1 Hills Road, Cambridge CB1 2EU, United Kingdom
[b] Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form 30 November 2013
Accepted 12 December 2013
Available online 8 January 2014

Keywords:
Differential item functioning
Effect size
Item response theory
Model fit
Polytomous items

ABSTRACT

Item bias or differential item functioning (DIF) has an important impact on the fairness of psychological and educational testing. In this paper, DIF is seen as a lack of fit to an item response (IRT) model. Inferences about the presence and importance of DIF require a process of so-called test purification where items with DIF are identified using statistical tests and DIF is modeled using group-specific item parameters. In the present study, DIF is identified using item-oriented Lagrange multiplier statistics. The first problem addressed is that the dependence of these statistics might cause problems in the presence of a relatively large number DIF items. Therefore, a stepwise procedure is proposed where DIF items are identified one or two at a time. Simulation studies are presented to illustrate the power and Type I error rate of the procedure. The second problem pertains to the importance of DIF, i.e., the effect size, and related problem of defining a stopping rule for the searching procedure for DIF. The estimate of the difference between the means and variances of the ability distributions of the studied groups of respondents is used as an effect size and the purification procedure is stopped when the change in this effect size becomes negligible.

1. Introduction

Differential item functioning (DIF) occurs when respondents with the same ability but from different groups (say, gender or ethnicity groups) have a different response probabilities on an item of a test or questionnaire [14]. Several statistical DIF detection methods have emerged in the last three decades [8,12,15,27,31,35,44,56,61,62,48,54]. During this period many researchers have reviewed various DIF detection methods (e.g., [9,28,40,47,55]. Most of the techniques proposed for the detection of DIF have been based on the evaluation of differences in response probabilities between groups conditional on some measure of ability. We can classify these techniques under two general categories: the first category is where a manifest score, such as the number-correct score, is taken as a proxy for ability and the second is where a latent ability variable of an IRT model functions as an ability measure.

The most common method used in the first category is the Mantel–Haenszel (MH) approach where DIF is evaluated by testing whether the response probability, given number-correct scores, differs between the groups. The MH test works quite well in practice under the Rasch model. Fischer [16,17], however, argues that its application under other IRT models raises several theoretical limitations. For instance, sufficient statistics are not available for the 2PL and 3PL models. Fischer's view on sufficient statistics equally applies to the log-linear approach where sum scores are used as proxies for ability; this view is also shared by Meredith and Millsap [38]. The observed score is nonlinearly related to the latent ability metric [14,35]

* Corresponding author. Tel.: +44 (0) 1223 558468.
E-mail addresses: Khalid.m@cambridgeesol.org (M.N. Khalid), c.a.w.glas@utwente.nl (C.A.W. Glas).

0263-2241/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2013.12.019

and factors such as guessing may preclude an adequate representation of the probability of correct response conditional on ability. Having said that, in general the correlation between the number-correct scores and ability estimates is quite high, so this is not the most important reason for considering alternative methods. The main problem arises in situations where the number-correct score loses its value as a proxy for ability. For example, there are test situations with large amounts of missing data and in the case of computer adaptive testing, where every student is administered a virtually unique set of items. In all these situations the number-correct score may not be appropriate for a meaningful assessment.

In an IRT model, ability is represented by latent variable θ, and a possible solution to the number correct score problem is to apply the MH and log-linear approach using subgroups that are homogenous with respect to an estimate of θ. This, however, introduces a different problem that the estimate of θ is subject to estimation error, which is difficult to take into account when forming the subgroups. An alternative is to view DIF as a special case of misfit of an IRT model and to use the machinery for IRT model-fit evaluation to explore DIF. An overview of this approach was given by Thissen et al. [63]. In that overview, evaluation of item parameter invariance over subgroups using Likelihood ratio and Wald statistics was presented as the main statistical tool for detection of DIF. Glas [20,21] argues that the Likelihood ratio and Wald approach are not very efficient because they require estimation of the parameters of the IRT model under the alternative hypothesis of DIF for every single item. To address these shortcomings, Glas [20,21] proposes using the Lagrange multiplier (LM) test by Aitchison and Silvey [1], and the equivalent efficient-score test [50], which do not require estimation of the parameters of the alternative model. Further, this approach supports the evaluation of many more model assumptions such as the form of the response function, unidimensionality and local stochastic independence, both at the level of items [24] and at the level of persons [23].

All methods listed above are seriously affected by the presence of high proportions of DIF items in a test and by the inclusion of DIF items in matching variable. To address this issue, several scale purification procedures have been suggested for the DIF detection methods, such as the two-stage or iterative Mantel–Haenszel method [27], the iterative Mantel method, the iterative generalized Mantel–Haenszel method [64,65], the iterative logistic regression method [19], and the iterative linking IRT-based method [7,46].

Scale purification procedures are useful in controlling Type I error rate and have high power when tests contain only a few DIF items. However, if tests have many DIF items, then DIF contamination cannot be completely eliminated by current scale purification procedures. Similar conclusions have been drawn when scale purification procedures were implemented on IRT-based DIF methods [7,34,46] and non-IRT-based DIF methods [10,19,26,27,39,45,64–66]. In this paper we propose an alternative scale purification method using Lagrange multiplier tests to address DIF contamination.

The significance of DIF, the extent to which the inferences made using test results are biased by DIF, is yet another important issue that needs to be looked at. The effect size of DIF is important to consider to avoid complicating inferences by practically trivial but statistically significant results. An example of a method to quantify the effect size is the DIF classification system for use with the MH statistical method developed by the Educational Testing Service [9,11]. In an IRT framework we propose to use an estimate of the difference between the means of the ability distributions of the studied groups of respondents as an effect size. This is motivated by the fact that ability distributions play an important role in most inferences made using IRT, such as in making pass/fail decisions, test equating, and the estimation of linear regression models on ability parameters as used in large scale education surveys such as NEAP, TIMSS and PISA.

In this paper we would first sketch a model of DIF and a concise framework of Lagrange multiplier test for the identification of DIF items. We would then present a number of simulation studies of the Type I error rate and power analysis. The difference between two versions of the LM test, one targeted at uniform DIF and one targeted at non-uniform DIF will be shown using a simulated example. This is followed by presenting an example using empirical data to show how the procedure works in practice. Finally, some conclusions are drawn, and suggestions for further research are provided.

2. Detection and modeling of DIF

In IRT models, the influences of items and persons on the observed responses are modeled by different sets of parameters. Since DIF is defined as the occurrence of differences in expected scores conditional on ability, IRT modeling seems especially fit for dealing with this problem. In practice, more than one DIF item may be present and therefore a stepwise procedure will be proposed where DIF items are identified one or two at a time. Both the significance of the test statistics and the impact of DIF are taken into account. The following procedure will be used here for detection and modeling of DIF. First, marginal maximum likelihood (MML) estimates of the item parameters and the means and variance parameters of the different groups of respondents are made using all items. Then an item is identified with the largest significant value on a Lagrange multiplier (LM) test statistic targeted at DIF. To model the DIF in this item, the item is given group-specific item parameters. That is, in the analysis, the item is split into two virtual items, one that is supposed to be given to the focal group and one that is supposed to be given to the reference group. Then, new MML estimates are made and the impact of DIF in terms of the change in the means and variances of the ability distributions is evaluated. If this change is considered substantial, the next item with DIF is searched for. The process is repeated until no more significant or relevant DIF is found. The assumptions of this procedure are that (1) the item which is mostly affected by DIF will have the largest value of the LM statistic regardless of the bias caused by the other items with DIF and (2) the change in the means and vari-