PSYCHOMETRIKA—VOL. 86, NO. 3, 712–716
SEPTEMBER 2021
https://doi.org/10.1007/s11336-021-09772-3

NOT ALL DIF IS SHAPED SIMILARLY

PAUL DE BOECK
THE OHIO STATE UNIVERSITY

SUN- JOO CHO
VANDERBILT UNIVERSITY

In response to the target article by Teresi et al. (2021), we explain why the article is useful and we also present a different approach. An alternative category of differential item functioning (DIF) is presented with a corresponding way of modeling DIF, based on random person and random item effects and explanatory covariates.

Key words: differential item functioning, explanatory covariates, random effects.

More than 10 years ago, Wainer (2010) wrote that the field has had “Enough, for Now” of differential item functioning (DIF) research in an invited article for the *Journal of Educational and Behavioral Statistics*. Despite this decree, DIF research has proven an enduring topic of research. Wainer explained that the two types of DIF detection methods, observed score methods such as Mantel–Haenszel and model-based likelihood ratio tests, are “more than enough to suit virtually any occasion” (p. 18). In his view, DIF methods research was no longer a priority. He seems to have underestimated the desire for further improvements. DIF methods apparently leave some room for improvement as discussed in the target article Teresi et.al (2021). The unstoppable continuation of DIF method research is focused on a traditional concept of DIF as a problem of individual items approached with a fixed effect method. In response to the focus on the traditional concept of DIF, we draw the attention here to another, perhaps more realistic, type of DIF. The alternative concept is one with random item and person effects and with covariates to explain the variation while allowing DIF to pervade subsets of items and even the whole test, if helpful to understand the item responses. At the same time, we agree with the authors of the target article that further method refinements for the more traditional concept are still important for purposes like the Patient-Reported Outcomes Measurement Information System (PROMIS) project and should not be stopped by decree. The refinements may be different for unidimensional and multidimensional tests, binary data and Likert scale data, and adaptive versus non-adaptive testing. There also are remaining issues for significance testing and determining effect size as discussed in the target manuscript.

The aim of our commentary is twofold. First, we will explain why and how the results of DIF detection methods may depend on the constraints imposed on the model (e.g., the use of anchor items). Second, in contrast with the target manuscript, our primary interest is in a different concept of DIF. The alternative concept is not primarily motivated by solving practical problems regarding DIF detection but rather by the need for a better understanding of item response data through modeling, which may help to solve DIF problems in the longer run.

It is well known that DIF models are not identified without constraints. In the case of uniform DIF, this is because of a trade-off between group mean differences and item

Correspondence should be made to Paul De Boeck, Department of Psychology, The Ohio State University, 240 Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43210, USA. Email: deboeck.2@osu.edu

difficulty DIF parameters while for non-uniform DIF the trade-off is between group variance ratios and item discrimination DIF parameters. The problem is the same for multidimensional models, except that multidimensional models are more complex because the covariance of the latent variables can play a role as illustrated in the target article. Different kinds of constraints for DIF detection have been used as described, for example, by Wang (2004) for the Rasch model. For non-uniform DIF, multiplicative equivalent constraints can be used. The most popular methods are anchor items, equal latent variable group means (no impact), equal mean item difficulties per group, and a fixed difference of group means based on a model estimation without DIF. The approaches are often accompanied by sequential testing (one item at a time) and a purification method. The score-based methods are also based on constraints, for example, the none-but-one DIF assumption.

Unless strong evidence exists for anchor items, the identification constraints are arbitrary. They lead to different but equally fitting results for the same data except when the constraints go beyond the minimum constraint necessary for model identification. The type of result and how much DIF inflation can be noticed depend on the combination of the constraints and the data. As illustrated in the target article, equal latent variable means leads to DIF inflation. How many DIF items are detected and how large the DIF magnitude seems to be depend on the data and the model constraints. These problems are alleviated when an iterative (purification) method is used. The implicit constraint of iterative procedures is a minimization of DIF, which can also be realized in a simultaneous way using regularization methods (e.g., Bauer et al., 2020; Magis et al., 2015; Tutz & Schauberger, 2015) as mentioned in the target article or outlier-based detection (Magis & De Boeck, 2011; Yuan et al., 2021). The identification issues can be avoided when uniform DIF is defined in terms of differences between differences: between-group differences in the within-group differences of the difficulty parameters (Bechger & Maris, 2015; Yuan et al., 2021), and when non-uniform DIF is defined in terms of ratios of ratios: between-group ratios of ratios of within-group discrimination parameters. Although this is a way to circumvent identification issues, it does not solve the issue of impact estimation.

While the minimization principle has clear practical value, it is not necessarily the best theoretical modeling option. Suppose that DIF pervades the whole test but only a few items are seriously affected. Identification and removal of these items as DIF items may seem to solve the problem, and perhaps it does for all practical purposes; however, if DIF is not an all-or-none concept (DIF vs. no DIF) but a graded concept instead, then it may yield a better understanding of the data not to use a minimization principle.

The alternative approach is DIF as a random variable across items or respondents and possibly across groups (e.g., Verhagen & Fox, 2013). It is rather unrealistic that a fixed DIF effect applies to all respondents of a group and that items are either DIF items or DIF-free items (i.e., that DIF is binary). A random effect approach can accommodate gradedness of DIF, and it also is a potential basis for using item covariates to explain DIF with the random effects functioning as residuals.

1. Random Item DIF

A full random item approach for DIF is proposed by De Boeck (2008). It implies a group effect and a random item variable for the difficulties in the reference group and the focal group, β_{ig} (i for items and g for group), with a covariance structure between groups, $\beta_{ig} \sim MVN(\mathbf{0}, \Sigma_{\beta})$. A model with only one random item variable can be compared with a model with a random item variable per group. This approach has led to the multiple item response profile model (Cho et al., 2012). An alternative model would be one with the same fixed item difficulties in the reference group and focal group plus a random item variable in the focal group to model DIF in that group.

When the model with an additional random item effect in the focal group is applied to the verbal aggression dataset (De Boeck & Wilson, 2004) with Gender as the group variable, it performs as well in terms of model selection criteria (e.g., AIC and BIC) as a model with fixed interaction effects between Gender and items. The additional random item effect for DIF has a standard deviation (SD) of 0.436 which is substantial compared with the standard deviation (SD) of 1.385 for the latent variable. An interesting advantage of the approach is that item covariates can be used to explain the random DIF variable and thus also the DIF (Van den Noortgate & De Boeck, 2005). For the same verbal aggression data, the SD of random DIF is reduced to 0.086 if the item covariate “mode” is used as an explanatory DIF item feature. In terms of the AIC and BIC, the explanatory model has a better fit than both the fixed interaction effect model and the random DIF effect model without an explanatory component. “Mode” refers to wanting vs. doing regarding verbal aggression, and it is clear from other analyses that the discrepancy between wanting and doing is larger for women than for men.

Such an explanatory success is not guaranteed because the source of DIF may be item specific (instead of shared by multiple items), in line with the traditional DIF concept. The random item DIF concept does not hinge on a shared source of DIF, but it is a handy approach that may answer the “why?” question regarding DIF, in line with Zumbo’s (2007) call for a next generation of DIF approaches.

Note that the random item DIF approach requires an identification constraint, which commonly is that the mean of the random item DIF variable in the focal group is zero. Though a nonzero mean could make sense, it has consequences for impact. Furthermore, random item approaches are still uncommon despite being parsimonious ways of modeling and handy for explanatory modeling.

2. Random Person DIF

When an item is biased in favor or disfavor of a group, it seems unlikely that it is to the same extent for all respondents of the group. Cho et al. (2016a) have proposed and described a secondary dimension model to handle individual differences in DIF for manifest groups (e.g., Gender). The model is a confirmatory model with a second dimension defined by the detected DIF items in the focal group. An important condition for this approach is that the detected DIF items share only one or just a few sources of differential functioning, possibly established with an explanatory approach as discussed for random item DIF. A similar method can be used for latent DIF, which is DIF detection for latent classes obtained from a mixture item response model (De Boeck et al., 2011; Cho et al., 2016b). This model has a secondary dimension in one mixture component and only one dimension in the other mixture component. Both of these secondary dimension models (with manifest groups or latent classes) can easily be extended with a secondary dimension for all respondents in the focal and reference group or in both latent classes. One limitation is that the models need to be confirmatory, meaning that a preliminary step would be required to derive a hypothesis Cheng et.al (2020), 2016b). Another limitation is that, when the DIF source is item specific, a multidimensional approach cannot help. That DIF may rely on multidimensionality is not new (e.g., Ackerman, 1992; Shealy & Stout, 1993), but an explicit modeling of the dimensions underlying DIF is still rare. We would rather not say that multidimensionality can “masquerade” as DIF (as in the target article) but rather that detecting DIF in the traditional way is an invitation to multidimensional modeling if the purpose is not just the identification of “bad” items. A clear advantage of the random person DIF is that the DIF can be isolated from the target dimension (Cho et al., 2016a, 2016b) without having to remove items and without the risk of distorted measurement.

A random DIF approach can possibly lead to a breakthrough regarding the explanation of impact. The minimization principle of DIF identification leads to a minimum number of DIF items, which may seem an advantage from a practical point of view, but allowing for a more pervasive and graded DIF may generate hypotheses about group differences that appear as impact but are in fact rooted in DIF. For example, it is possible that the items are rather homogeneous in terms of an item covariate that does cause DIF, which shows in a small variance of random item DIF. Because of the homogeneity of the items (e.g., all having a high value on the item covariate), it is possible that the DIF is not noticed and shows as impact instead (because all items have a similarly high value on the item covariate). Larger differences in the item covariate across items could reveal the DIF and its source.

In summary, although refinements for the detection of DIF as defined in the traditional way are important as illustrated in the target article, we also see potential in alternative DIF concepts for the purpose of a better understanding of item responses and possibly, in the longer run, for dealing with DIF based on this better understanding.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x.

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 43–55. https://doi.org/10.1080/10705511.2019.1642754.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*, 317–340. https://doi.org/10.1007/s11336-014-9408-y.

Cheng, C.-P., Chen, C.-C., Shin, C.-L. (2020).An exploratory strategy to identify and define sources of differential item functioning. *Applied Psychological Measurement, 44*, 548-560. https://doi.org/10.1177/0146621620931190.

Cho, S.-J., Partchev, I., & De Boeck, P. (2012). Parameter estimation of multiple item profiles models. *British Journal of Mathematical and Statistical Psychology*, *65*, 438–466. https://doi.org/10.1111/j.2044-8317.2011.02036.x.

Cho, S.-J., Suh, Y., & Lee, W.-Y. (2016a). After DIF items are detected: IRT calibration and scoring in the presence of DIF. *Applied Psychological Measurement*, *40*, 573–591. https://doi.org/10.1177/0146621616664304.

Cho, S.-J., Suh, Y., & Lee, W.-Y. (2016b). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*, 48–61.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. https://doi.org/10.1007/s11336-008-9092-x.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*, 583–603. https://doi.org/10.1177/0146621611428446.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Magis, D., & De Boeck, P. (2011). A robust outlier approach to prevent Type 1 error inflation in DIF. *Educational and Psychological Measurement*, *72*, 291–311. https://doi.org/10.1177/0013164411416975.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the Lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135. https://doi.org/10.3102/1076998614559747.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale: Lawrence Erlbaum.

Teresi, J. A., Wang, C., Kleinman, M., Jones, B. N., & Weiss, D. J. (2021). Differential item functioning analyses of the patient reported outcomes measurement information system (PROMIS) measures: Methods, challenges, advances, and future directions. *Psychometrika*.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*, 21–43. https://doi.org/10.1007/s11336-013-9377-6.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464. https://doi.org/10.3102/10769986030004443.

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383–401.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, *35*, 5–25. https://doi.org/10.3102/1076998609355124.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221–261. https://doi.org/10.3200/JEXE.72.3.221-261.

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*. https://doi.org/10.1007/s11336-021-09746-5.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233. https://doi.org/10.1080/15434300701375832.

Manuscript Received: 18 APR 2021
Final Version Received: 18 APR 2021
Published Online Date: 5 JUN 2021