Promise and peril: Agnostic identification methods for detecting differential item functioning

Ben Stenhaug, Ben Domingue, and Mike Frank
Stanford University

Abstract

It is well known that likelihood ratio tests (LRT) are effective at detecting differential item functioning (DIF) in item response models. However, to use an LRT, the model needs to be identified so that differences in group ability can be disentangled from potential DIF. We summarize existing agnostic identification (AI) methods and propose a variety of new methods. We conduct a simulation study — which we believe to be more realistic than most DIF simulation studies in the literature — and find that one of the proposed new AI methods, All-others-as-anchors-one-at-a-time (AOAA-OAT), significantly outperforms current methods. We also offer a new method, the equal means, multiple imputation logit graph (EM-MILG), that presents clearly all information about possible DIF, including sampling variability in item parameters, to the analyst.

Contents

Table 1: Summary of agnostic identification methods

Method	Description	Literature
AOAA	Test if each item has DIF by using all of the other items as anchors (not iterative).	Originally proposed by Lord (1980) and formalized by Thissen et al. (1993)
AOAA-AS	The first iteration is AOAA. All items that test positive for DIF are removed from the anchor set. Continue iterating until no new items test positive for DIF.	Proposed by Drasgow (1987)
AOAA-OAT	The first iteration is AOAA. Only the item that shows the most extreme DIF is removed from the anchor set. Continue iterating until no new items test positive for DIF.	To our knowledge, not proposed or used previously
equal means clustering (EMC)	Cluster items based on differences in performance across groups and choose one of the clusters to be the anchor cluster.	Proposed by Bechger and Maris (2015) and refined by Pohl et al. (2017)
equal means, multiple imputation logit graph (EM-MILG)	Arbitrarily set both group means to 0, which pushes all group performance differences to the item paramaters, measure variability using multiple imputations, and graph the result. Can be used by an analyst to hand select anchor items	Inspired by pedagolical examples Pohl et al. (2017) and Talbot III (2013)
multiple imputation logit graph (MILG)	Similar to EM-MILG but used to visualize potential DIF once the model is already identified	
maximizing Gini index (MAXGI)	Arbitrarily set both group means to 0 and then choose an anchor point by maximizing the Gini index	Adapted from work by Strobl et al. (2018)
minimizing the area between curves (MINBC)	Of the infinite number of model that maximizes the likelihood of the data, choose the one with the minimum total area between the two groups' item characteristic curves	Built on and inspired by work by Raju (1988) and more recently, Chalmers et al. (2016)

References

Bechger, T. M. and Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, 80(2):317–340.

Chalmers, R. P., Counsell, A., and Flora, D. B. (2016). It might not make a big dif: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76(1):114–140.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied psychology*, 72(1):19.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Pohl, S., Stets, E., and Carstensen, C. H. (2017). Cluster-based anchor item identification and selection.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4):495–502.

Strobl, C., Kopf, J., Hartmann, R., and Zeileis, A. (2018). Anchor point selection: An approach for anchoring without anchor items. Technical report, Working Papers in Economics and Statistics.

Talbot III, R. M. (2013). Taking an item-level approach to measuring change with the force and motion conceptual evaluation: An application of item response theory. *School Science and Mathematics*, 113(7):356–365.

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.