The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2017

Volume 8, Pages 103-106

IConSE 2017: International Conference on Science and Education (IConSE)

DIF ANALYSES WITH MANIFEST AND LATENT GROUPS: ANALYSES OF PISA 2012 MATHEMATICS DATA FROM TURKEY

Tugba Karadavut
Kilis 7 Aralık University

Abstract: Differential item functioning (DIF) indicates existence of items in a test on which different groups of examinees perform differentially. The groups in DIF analyses are typically designated based on their manifest characteristics such as gender and ethnicity. Previous research showed that, examinees of a manifest group may not be homogeneous on the dimension that is actually causing DIF. That is, the manifest groups have a weak relationship with the latent groups that explicit true differential performance on items. In this study, DIF items on the basis of gender were identified for PISA 2012 mathematics data from Turkish subsample. Then, latent groups in the subsample were estimated in order to detect the true groups that perform differentially.

Keywords: Differential item functioning, DIF, item response theory, IRT, likelihood ratio test, manifest DIF, latent DIF

Introduction

Unidimensional item response theory (IRT) models assume that the underlying latent trait that explains the common variance among item responses is one dimensional (Lord & Novick, 1968). In other words, the items are uncorrelated for fixed values of the underlying latent trait (i.e., local independence; McDonald, 1999), because no additional dimension is sufficiently dominant to explain a substantial amount (i.e., 20% or higher) of the common variance among items (Reckase, 1979)[1].

Differential item functioning (DIF) (e.g., Holland & Wainer, 1993) refers to performance differences between groups on certain items after the groups have been matched on the latent ability that is intended to be measured by the item (Dorans & Holland, 1993). In other words, the groups perform differentially on these items not because they differ on the latent ability that is intended to be measured, but because they differ on a nuisance dimension which is not of interest (Ackerman, 1992). Differential item functioning threaten construct validity in the assesment because of unintended multidimensionality in the construct that is being measured (Steinberg & Thissen, 1996). DIF analyses ensure that interpretations of test scores are valid for all distinct groups of the examinees (Zwick, 2012).

The groups in DIF analyses are typically determined based on their manifest characteristics such as gender, race and ethnicity. There are two issues worth of consideration regarding this approach. First, the manifest groups are not homogeneous on the dimension that is actually causing DIF (Samuelsen, 2005). Second, the typical approach of DIF detection identifies the items with DIF, however does not explain the the dimension that is actually causing DIF (Cohen & Bolt, 2005).

Mixture item response theory models can be used to identify the latent groups that actually perform differentially on items (Cohen & Bolt, 2005). DIF analysis that is based on manifest groups could be called *manifest* DIF and DIF analysis that is based on latent groups could be called *latent DIF* (Cho, Suh, & Lee, 2016). The manifest groups in manifest DIF and the latent groups that are detected in latent DIF are often not comparable (Cohen & Bolt, 2005).

In this study, DIF items based on gender were identified for PISA 2012 mathematics data from Turkey. Then, latent groups in Turkish data were estimated by using a mixture 2-parameter logistic (2PL) IRT model in order

- Selection and peer-review under responsibility of the Organizing Committee of the conference

*Corresponding author: Tugba Karadavut-E-mail: tugbakaradavut@kilis.edu.tr to detect the true groups that perform differentially on the items as well as to determine the items that indicate DIF.

Methods

Manifest DIF

There are different methods and approaches for DIF detection (e.g., the Mantel–Haenszel procedure, the Standardization procedure, logistic regression, logistic discriminant function analysis, Lord's chi-square, Raju's area measures, likelihood ratio test). The IRT model-based likelihood ratio test (LR) was used in this study for detection of DIF (Thissen, Steinberg, & Gerrard, 1968; Thissen, Steinberg, & Wainer, 1993). IRTLRDIF (Thissen, 2001) software was used to conduct the LR analysis.

Latent DIF

A mixture 2PL IRT model was used to detect the true groups that perform differentially. A 2PL IRT model is one of the unidimensional IRT models that is for dichotomous items (e.g., multiple choice). The 2PL model defines the probability that an examinee j with ability θ answers item i correctly ($P_i(\theta_j)$) by the following equation:

$$P_i(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}, \tag{1}$$

where b_i is the item difficulty parameter for item i and a_i is the item discrimination parameter for item i.

A mixture 2PL IRT model defines the probability of a correct response to item *i* by examinee *j* as:

$$p_{ji} = P(X_{ij} = 1 | \theta_{jg}) = \sum_g \pi_g \frac{\exp[\mathrm{a}_{ig}(\theta_{jg} - b_{ig})]}{1 + \exp[\mathrm{a}_{ig}(\theta_{jg} - b_{ig})]}, \tag{2}$$

where θ_{jg} is the examinee's ability in latent group g, and b_{ig} is the item difficulty parameter in latent group g, and a_{ig} is the item discrimination parameter in latent group g.

Estimation of mixture 2PL IRTmodel parameters

Estimation of parameters in the mixture 2PL IRT model was done by using the Markov Chain Monte Carlo (MCMC) method as implemented in the computer software OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). A burn-in period of 3,000 iterations was used with a total number of 15,000 iterations. Following priors were used for MCMC estimation of model parameters:

$$\begin{gathered} b_i \sim \text{Normal}(0,1), \qquad i = 1, \ldots, n, \\ a_i \sim \text{Normal}(0,1) \text{ and } a_i > 0\,, \; i = 1, \ldots, n, \\ \theta_j \sim \text{Normal}(0,1), \qquad j = 1, \ldots, N. \end{gathered} \tag{3}$$

Dataset

The data used in this study is from the 2012 cycle of the Program for International Student Assessment (PISA; OECD, 2014). In this example, data from Turkey was analyzed with a sample size of 351. Data from booklet 5 was used resulting 36 items to be analyzed. Examinees were deleted from the dataset if they provided a missing response to at least one item. The partial credit items were recoded dichotomously. That is, a full credit was recoded as a correct response and a partial credit was recoded as an incorrect response.

Results and Findings

LR analysis detected seven items with DIF based on gender. Mixture 2PL IRT analysis resulted in one underlying latent class based on both Akaike's information criterion (AIC; Akaike, 1974) and based on Schwarz's Bayesian information criterion (BIC; Schwarz, 1978).

Conclusion

Mixture 2PL IRT model yielded only one underlying latent class. That is, there were not any groups that perform differentially on the test items. The LR DIF analyses on the other hand yielded that female and male students performed differentially on seven items. Detection of groups that perform differentially seems to be misleading when the grouping is done based on gender. It is because no underlying latent groups detected that performs differentially on the given items.

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

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

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. British Journal of Mathematical and Statistical Psychology, 63, 395-416.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine, 28*, 3049-3082.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

OECD. (2014), *PISA 2012 Technical Report*. Retrieved October 1, 2017, from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement*, *73*, 5-26.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland, College Park.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81-97.

Thissen, D. (2001). Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill: University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report. No. RR-12-08). Princeton, NJ: Educational Testing Service.

Footnotes

[1] Local independence and unidimensionality are related yet different concepts. However, a multidimensional IRT model and a locally dependent IRT model are indistinguishable in practice (Ip, 2010), endorsing that the local dependency may be an indicator of multidimensionality (Reise, Scheines, Widaman, & Haviland, 2013; Steinberg & Thissen, 1996).