Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
April 2025, Vol. 00, No. 0, pp. 1–3

Digital Module 38: Differential Item Functioning by Multiple Variables Using Moderated Nonlinear Factor Analysis

Sanford R. Student and Ethan M. McCormick

Module Abstract: *When investigating potential bias in educational test items via differential item functioning (DIF) analysis, researchers have historically been limited to comparing two groups of students at a time. The recent introduction of Moderated Nonlinear Factor Analysis (MNLFA) generalizes Item Response Theory models to extend the assessment of DIF to an arbitrary number of background variables. This facilitates more complex analyses such as DIF across more than two groups (e.g. low/middle/high socioeconomic status), across more than one background variable (e.g. DIF by race/ethnicity and gender), across non-categorical background variables (e.g. DIF by parental income), and more. Framing MNLFA as a generalization of the two-parameter logistic IRT model, we introduce the model with an emphasis on the parameters representing DIF versus impact; describe the current state of the art for estimating MNLFA models; and illustrate the application of MNLFA in a scenario where one wants to test for DIF across two background variables at once.*

Keywords: differential item functioning, factor analysis, item response theory, measurement invariance, moderated nonlinear factor analysis

Differential item functioning (DIF) analysis has a rich history in the educational measurement field (e.g., Angoff & Ford, 1973; Holland & Thayer, 1986; Swaminathan & Rogers, 1990; Thissen, 2024). DIF analysis is the investigation of the extent to which responses to a given item differ systematically according to a background variable of interest (e.g., race/ethnicity, socioeconomic status, gender), *above and beyond* differences in item responding implied by group-to-group differences on the construct of interest, often called θ for the purposes of modeling. DIF in educational measurement is generally cast as an important, though certainly noncomprehensive, threat to the fairness and validity of an assessment (AERA et al., 2014), so its assessment is a key step in nearly all test development.

Most mainstream DIF analysis methods, especially those in the Item Response Theory (IRT) paradigm, are group-based and consider a small number of groups at a time, often just two (a “focal” and a “reference” group). This multiple-group approach to DIF is an inherently limited analytic framework: (1) analysis of DIF by continuous background variables in most frameworks requires coarsening the variable into a categorical one, thereby losing information, and (2) analysis by multiple background variables is generally not possible, even though there is good reason to consider the intersection of different variables in the context of DIF analysis, as enacted in a recent ITEMS Module (Russell, 2024). This can be partially, but not fully, addressed by models typically used for measurement invariance analysis in structural equation modeling (SEM): while multiple indicator, multiple cause (MIMIC) and MIMIC-with-interaction models (Jöreskog & Goldberger, 1975; Woods & Grimm, 2011) facilitate analysis of DIF by multiple variables, their limitation is the assumption of constant variance across all background variables. This is one advantage of multiple-group models: they do allow the mean *and variance* of θ to differ by group. That is, the features needed to conduct principled analysis of DIF by multiple background variables at once have historically been split across two distinct analytic frameworks.

Moderated Nonlinear Factor Analysis (Bauer, 2017; Bauer & Hussong, 2009) is a general measurement modeling framework with foundations in both IRT and SEM. For the purposes of DIF analysis by multiple background variables, MNLFA incorporates the best features of both multigroup IRT and MIMIC approaches. Like MIMIC models, MNLFA allows for an arbitrary number of background variables (either continuous or categorical). Like multigroup IRT models, MNLFA freely estimates the variance of the θ distribution as a function of variables being analyzed for DIF. Yet, the flexibility of MNLFA also leads to model identification complexities and challenges for estimation. This ITEMS module walks the user through the conceptual foundations of DIF analysis by an arbitrary number of background variables using MNLFA, and describes how penalized maximum likelihood estimation can be used to reduce the complexity of models with many DIF parameters (Bauer et al., 2020; Belzak & Bauer, 2024). Users completing this module will go forward equipped with a powerful new approach to DIF analysis whose flexibility enables analyses that are simply not possible using traditional DIF methods (Bauer, 2023).

Learning Objectives

Upon completion of this ITEMS module, learners should be able to:

- Articulate the difference between uniform and nonuniform DIF in the slope-intercept form of the 2PL IRT model.
- Differentiate DIF from impact, and describe the implications of both for parameters of traditional IRT models.
- Describe how moderated nonlinear factor analysis can be applied to estimate both DIF and impact in the slope-intercept 2PL.
- Apply regularized moderated nonlinear factor analysis to simultaneously estimate DIF and impact for multiple covariates of mixed types (i.e., categorical and continuous) using the R package *regDIF*.
- Use the results of regularized MNLFA estimation to inform next steps in DIF analysis.

Brief Description of Each Section

The digital ITEMS module is divided into the following sections, which can be reviewed sequentially or independently.

- **Section 1—IRT, DIF, and a More General Model**: This section provides a brief refresher on the general concepts of DIF and impact. We review the two-parameter logistic IRT model, demonstrating its isomorphism with a common categorical factor analysis model that uses a slope and intercept instead of a discrimination and difficulty. In the context of this model, we introduce uniform and nonuniform DIF in terms of their implications for item response functions. We review the affordances and limitations of multiple-group and MIMIC approaches to DIF analysis, introducing MNLFA as a "best of both worlds" unifying framework.
- **Section 2—An Overview of MNLFA for DIF and Impact Assessment**: This section provides a detailed conceptual overview of how MNLFA models DIF and impact. Using model equations, path diagrams, item response function curves, and graphical depictions of model-implied θ distributions, we outline the four major relations that MNLFA incorporates into the 2PL IRT model: impact on the mean of θ, impact on the variance of θ, moderation of item intercepts (uniform DIF), and moderation of item slopes (nonuniform DIF).
- **Section 3—MNLFA Estimation and Interpretation**: This section carries the conceptual summary from Section 2 into an example analysis of simulated data. We first describe the challenges of estimating MNLFA models, noting regularized MNLFA as an approach that gets around some of these challenges, particularly the complexity of modeling many DIF parameters when one generally expects few items to exhibit DIF. We provide further guidance on interpreting hypothetical values for estimates of model parameters representing both DIF and impact.
- **Section 4—MNLFA Applied Example**: Here, we walk through regularized MNLFA analysis of the simulated dataset in more detail. We use the *regDIF* R package (Belzak, 2023) to conduct LASSO regularized estimation of an MNLFA including DIF and impact by two background variables.
- **Section 5—MNLFA Code Walkthrough**: We guide the user through preparing data and estimating DIF models. We outline the wealth of information provided by *regDIF*, focusing on the most salient results for DIF analysis. We also give an overview of software options for estimation of MNLFAs more broadly using maximum likelihood or Markov Chain Monte Carlo (Bauer, 2017; Brandt et al., 2023; Chen et al., 2022; Enders et al., 2024; Kolbe et al., 2024).
- **Guided Activity**: We provide the simulated data used in this module and a guide outlining the steps to conduct regularized MNLFA analysis of the dataset. We include a worked example that reproduces all of the results presented in the module. This example also extends the analysis to re-estimate the model via standard maximum likelihood in the software Mplus (Muthén & Muthén, 2017), as suggested by Bauer et al. (2020).

In the portal site, you will find a video version of the core content as well as a handout with our slides, guided activity, list of further reading, and dataset.

Audience

The audience for this module ranges from graduate students in quantitative methods and evaluation, to faculty teaching educational measurement or related quantitative methods courses, to practitioners conducting DIF analysis for operational tests. Graduate students with some prior exposure to IRT and DIF—for example, via a first semester course in measurement—will benefit from learning about a uniquely flexible approach to DIF analysis that subsumes several previously distinct frameworks to overcome the limitations of each. We hope that students will carry this learning forward into their own research and practice as they enter the field. Faculty members teaching measurement and psychometrics will gain an additional resource for teaching graduate students about different approaches to conducting DIF analysis, while instructors of advanced SEM courses may find it interesting to contrast the use of MNLFA for DIF analysis enacted in this module with the use of MNLFA for measurement invariance testing in SEM (Kolbe et al., 2024). Finally, we hope that psychometricians working in operational settings will be interested in applying these methods in their own work, given the potential limitations of assessing DIF "two groups at a time" or "one variable at a time."

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B240013 to the University of Delaware. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank NCME ITEMS Editor Stella Kim for her thorough and consistent support in the completion of this module. Thank you to Gregory R. Hancock, Daniel J. Bauer, and James G. Soland for their helpful peer review.

ORCID

Sanford R. Student https://orcid.org/0000-0001-7419-2437
Ethan M. McCormick https://orcid.org/0000-0002-7919-4340

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(95–106), p. 13. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), pp. 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology*, *76*(3), pp. 435–461. https://doi.org/10.1111/bmsp.12316

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*(1), pp. 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), pp. 101–125. https://doi.org/10.1037/a0015583

Belzak, W. C. M. (2023). The regDIF R package: Evaluating complex sources of measurement bias using regularized differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), pp. 974–984. https://doi.org/10.1080/10705511.2023.2170235

Belzak, W. C. M., & Bauer, D. J. (2024). Using regularization to identify measurement bias across multiple background characteristics: A penalized expectation–maximization algorithm. Journal of Educational and Behavioral Statistics, Advance online publication. https://doi.org/10.3102/10769986231226439

Brandt, H., Chen, S. M., & Bauer, D. J. (2023). Bayesian penalty methods for evaluating measurement invariance in moderated nonlinear factor analysis. Psychological Methods, Advance online publication. https://doi.org/10.1037/met0000552

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), pp. 122–139. https://doi.org/10.1080/10705511.2021.1948335

Enders, C. K., Vera, J. D., Keller, B. T., Lenartowicz, A., & Loo, S. K. (2024). Building a simpler moderated nonlinear factor analysis model with Markov Chain Monte Carlo estimation. *Psychological Methods*, Online ahead of print. https://doi.org/10.1037/met0000712

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), pp. 1–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), p. 631. https://doi.org/10.2307/2285946

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2024). Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, *29*(2), pp. 388–406. https://doi.org/10.1037/met0000501

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical analysis with latent variables: User's guide* (Version 8).

Russell, M. (2024). Digital module 36: Applying intersectionality theory to educational measurement. *Educational Measurement: Issues and Practice*, *43*(3), pp. 106–108. https://doi.org/10.1111/emip.12622

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), pp. 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D. (2024). A review of some of the history of factorial invariance and differential item functioning. Multivariate Behavioral Research, Advance online publication. https://doi.org/10.1080/00273171.2024.2396148

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with Multiple Indicator Multiple Cause models. *Applied Psychological Measurement*, *35*(5), pp. 339–361. https://doi.org/10.1177/0146621611405984