Using continuous background variables in DIPF detection

Lela Roos, student ID: 1729683,
FECT: 22-1892, word count: 2497

January 2023

1 Introduction

In many educational settings the interest lies in comparing groups on a particular measure. For example, we want to see whether boys or girls have a higher math ability. In order to make valid statements about the (mean) ability of boys and girls, we need to ensure that the measure or test is comparable across both groups. When the conditions hold such that scores are comparable across groups for a particular measurement model, this is called *measurement invariance* (Millsap, 2011). When a test is not measurement invariant, at least one item in a test functions differently across groups. This is referred to as *differential item functioning* (DIF).

Mellenbergh defined an item to be unbiased (i.e. no DIF) when the distribution of the item responses only depends on the latent trait scores, and not on the combination of latent trait and the grouping variable (Mellenbergh, 1989). In the case of dichotomously scored items, and a nominal grouping variable, this can also be expressed as the requirement that the probability of a given response is the same for members of different groups with the same position on the trait measured by the test (Mellenbergh, 1989):

$$P(Y = 1|x, \eta) = P(Y = 1|\eta) \text{ for all } x \text{ and } \eta, \tag{1}$$

where Y is the item response, x is a grouping variable, and η [1] is the latent ability variable. Two theoretical frameworks from which measurement invariance and DIF are usually addressed are Structural Equation Modelling (SEM) and Item Response Theory (IRT), respectively.

Starting with IRT models, which are statistical models that describe the mathematical relationship between item and person parameters (Hambleton et al., 1991). IRT models can include a different number of item parameters that, for example, tell us something about the location, discriminating ability, guessing and slipping impact of the item (Janssen, 2018). The 1-parameter logistic model (1PLM) only includes the item location or difficulty parameter as an item parameter, it can be expressed as follows:

$$P(Y_{ip} = 1|\eta_p) = \frac{e^{\eta_p - \delta_i}}{1 + e^{\eta_p - \delta_i}}, \tag{2}$$

where Y_{ip} is the item response for person on item i, and δ_i is the item difficulty of item i. The 2-parameter logistic model (2PLM) is an extension of the 1PLM that also includes an item discrimination parameter, which represents how well an item is able to differentiate among persons (Janssen, 2018):

$$P(Y_{ip} = 1|\eta_p) = \frac{e^{\alpha_i(\eta_p - \delta_i)}}{1 + e^{\alpha_i(\eta_p - \delta_i)}}, \tag{3}$$

where Y_{ip} and δ_i are defined as before and α_i is item i's discrimination parameter. DIF detection comes down to checking whether model parameters are the same across different groups (Janssen, 2018). Different methods to compare parameters exist (e.g. Lord, 1976; Raju, 1988; Swaminathan and Rogers, 1990). There are a few issues with these commonly used IRT DIF detection methods.

[1] Latent ability is often represented as θ within Item Response Theory literature (see e.g. Janssen, 2018; Mellenbergh, 1989), here I use η which is the convention within Structural Equation Modeling literature (see e.g. Bauer and Hussong, 2009; Meredith, 1993).

Cuellar Caicedo (2022) highlights the key issues: the circularity problem and the identification problem. The circularity problem refers to the need for establishing non-DIF items before DIF detection. Non-DIF items are needed to be able to correctly estimate latent ability scores, and we condition on this estimated ability when trying to detect bias (Mellenbergh, 1989). We, however, often do not know which items are DIF-free. Furthermore, the identification problem highlights that individual item characteristics are not identified. We can thus not analyze a particular item's functioning across group. However, as Bechger and Maris (2014) address, we can investigate the relative positioning of an item. Based on this idea, they introduced the concept of differential item-*pair* functioning (DIPF). Their work was limited to comparing two groups, and Cuellar Caicedo (2022) extends this technique to multiple groups.

Now turning to SEM models, which can be seen as combinations of factor analysis and path analysis (Weston & Gore, 2006). One well-known SEM model is the Confirmatory Factor Analysis (CFA) model, which can be expressed as folows:

$$Y_{ip} = \nu_i + \lambda_i \eta_p + \epsilon_p, \tag{4}$$

where Y_{ip} is the observed score for person p on item i, and ν_i and λ_i are the item-specific intercept and factor loading, respectively. The errors are assumed to be normally distributed: $\epsilon_p \sim N(0, \Theta)$.

Within SEM, testing for measurement invariance is often done using Multi-Group Confirmatory Factor Analysis (MGCFA). This means a CFA model is fit in all groups, after which equality constraints are placed on the model. Different levels of invariance have been defined (Meredith, 1993). Testing for these different forms of measurement invariance with MGCFA is done in a step-wise procedure (Millsap, 2011; van de Schoot et al., 2012). First, it is checked if the model structure is the same across groups (*configural invariance*). Then, the following parameters are constrained to be equal (in this order): factor loadings (*metric invariance*), intercepts (*scalar invariance*), residual variances (*strict factorial invariance*).

These forms of invariance can also be related to the DIF terminology. DIF is usually split up into uniform and non-uniform DIF. Uniform DIF refers to the scenario where, for a DIF item, the relationship between the item response and background variable is the same across the levels of the latent trait (Glas & Ouborg, 1993). Uniform DIF can thus be seen as a violation of scalar invariance. Non-uniform DIF refers to the case where the relationship between the item response and background variable differs across levels of the latent trait (Glas & Ouborg, 1993), which would be a violation of metric invariance.

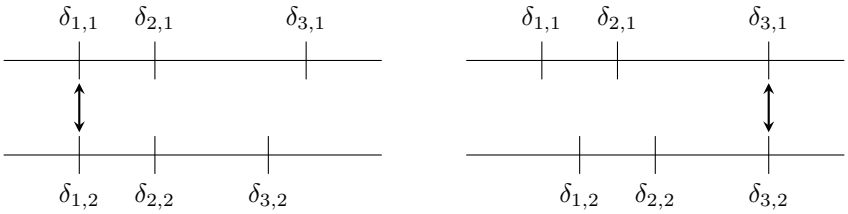
Fitting a MGCFA with many groups can become an unmanageable task because of the large amount of estimated parameters (Asparouhov & Muthén, 2014). Splitting the data and fitting a CFA in each group can also lead to problems of low power (Kolbe, 2022). Single-group methods, where the data is not split up, form an alternative. Kolbe (2022) discussed multiple single-group methods, her analyses show that Moderated Nonlinear Factor Analysis (MNLFA) performs well under many different conditions and allows for continuous background variables.

So, traditional DIF and measurement invariance analyses are often focused on comparing individuals on one dichotomous grouping variable. More recently, developments have been made to be able to compare across multiple groups (e.g. Cuellar Caicedo, 2022; Kim et al., 2017). However, these methods are still limited to testing across a categorical variable. This thesis is focused on extending analyses to include continuous background variables. I aim to bring recent developments within the IRT and SEM frameworks together: Bechger and Maris (2014)'s DIPF method and MNLFA. The concept of using a continuous background variable as a moderator from MNLFA could provide another extension of the DIPF method that allows for continuous background variables. This leads to the following research question: *How can continuous background variables be used in DIPF-detection?*

2 Theoretical background

2.1 Differential Item-*Pair* Functioning

Bechger and Maris (2014)'s concept of differential item-*pair* functioning (DIPF) is based on the fact that the item difficulties of the 1PLM (and other IRT models) are not identified. Models are identified using a normalization, estimates of the item difficulties are dependent on this normalization. Since item difficulties are not identified parameters, DIF can also not be defined as



(a) $\delta_{1,1} = \delta_{1,2} = 0$ (b) $\delta_{3,1} = \delta_{3,2} = 0$

Figure 1: Item difficulties are shown for three items in two groups, where $\delta_{i,g}$ is the difficulty parameter for item i in group g, under two different normalizations.

a property of an item. When there is DIF, the normalization can affect which items are flagged as DIF items. This can be seen in figure 1, which shows the item difficulties for three items in two groups. The relative difficulties between item 3 and the other two items differ per group. In figure 1a, item 3 would be considered as exhibiting DIF. However, in figure 1b, the other two items would be flagged as DIF items. Bechger and Maris (2014)'s DIPF method is based on the fact that relative item difficulties are identified. As can be seen from figure 1, our conclusions about relative item difficulties would be the same regardless of the normalization (namely, that only the relative difficulty between item 1 and 2 is the same across the two groups).

Instead of DIF we thus consider DIPF, an item-pair shows DIPF when its relative difficulty is not consistent across a background variable. Bechger and Maris (2014) focus their discussion on DIPF detection when comparing two groups. Then, the null hypothesis that is tested is:

$$H_0 : \delta_{ij}^{(1)} = \delta_{ij}^{(2)}, \tag{5}$$

where $\delta_{ij}^{(g)}$ is the relative difficulty between item i and item j in group g, i.e. $\delta_{ij}^{(g)} = \delta_i^{(g)} - \delta_j^{(g)}$. The statistic to test DIPF for item-pair i and j is given by:

$$\hat{D}_{ij} = \frac{\hat{\delta}_{ij}^{(1)} - \hat{\delta}_{ij}^{(2)}}{\sqrt{Var(\hat{\delta}_{ij}^{(1)}) + Var(\hat{\delta}_{ij}^{(2)})}}. \tag{6}$$

Under the null hypothesis, this standardized difference $\hat{D}_{ij}$ is asymptotically standard normal, which allows for straightforward testing procedures. For example, with a significance level α of 0.05, H_0 is rejected when $|\hat{D}_{ij}| > 1.96$. Bechger and Maris (2014) also developed a test to simultaneously identify DIPF across all item-pairs in a test. The relative difficulties of all other items in comparison to one reference item contain all needed information for this test (Bechger & Maris, 2014; Cuellar Caicedo, 2022). Cuellar Caicedo (2022) uses the term *relative difficulty profile* (RDP) to refer to the set of relative difficulties for a reference item. When comparing two groups, the null hypothesis for the test for DIPF across the full set of items then becomes (Bechger & Maris, 2014):

$$H_0 : RDP^{(1)} - RDP^{(2)} = 0, \tag{7}$$

where $RDP^{(g)}$ is the RDP for group g.

Bechger and Maris (2014) show that the following squared Mahalanobis distance can be used to test this hypothesis:

$$\chi^2_\beta = \hat{\beta}^T \Sigma^{-1} \hat{\beta} \tag{8}$$

where $\hat{\beta} = \hat{RDP}^{(1)} - \hat{RDP}^{(2)}$. When H_0 holds, the statistic in Equation 8 follows a chi-squared distribution with $k-1$ degrees of freedom (where k is the number of items). See Cuellar Caicedo (2022) for a discussion on how to extend these DIPF tests to testing invariance across more than two groups.

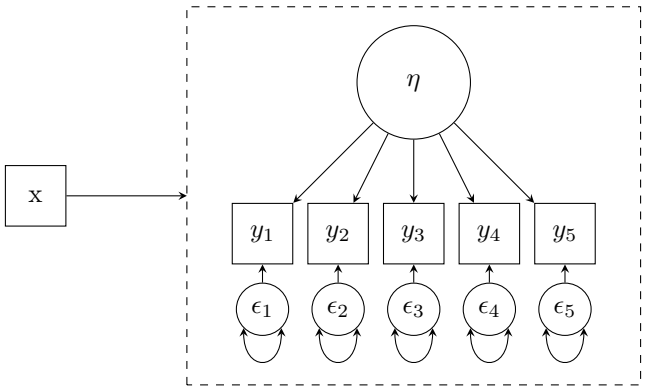


Figure 2: A graphical representation of a simple MNLFA model. η represents a latent factor, with 5 indicators y_1 to y_5, and their measurement errors ϵ_1 to ϵ_5. The background variable x can have an effect on all model parameters within the dashed rectangle.

2.2 Moderated Nonlinear Factor Analysis

Before turning to the discussion of the MNLFA model, I first discuss the Generalized Linear Factor Analysis (GLFA) Model because the MNLFA model can be seen as an extension of the GLFA model. Inspired by the General Linear Model (GLM), the GLFA can be expressed as follows:

$$g_i(\mu_{ip}) = \nu_i + \lambda_i \eta_p, \tag{9}$$

where $g_i(\cdot)$ is a link function, μ_{ip} is the expected value of item i for person p, ν_i and λ_i are the intercept and factor loading for item i, and η_p is the latent factor score for person p, for which a normal distribution is assumed such that $\eta_p \sim N(\alpha, \phi)$.

Bauer and Hussong (2009) and Kush et al. (2022) highlight how the GLFA can be related to the 2PLM. To make the GLFA equivalent to the 2PLM, a Bernoulli response distribution for each item is chosen, such that $y_{ip}|\eta_p \sim Ber(\mu_{ip})$. Additionally, the logit link function defined as $g_i(\mu_{ip}) = \ln[\mu_{ip}/(1-\mu_{ip})]$ is used. Thus, we have:

$$\ln(\frac{\mu_{ip}}{1-\mu_{ip}}) = \nu_i + \lambda_i \eta_p, \tag{10}$$

Using the inverse logit link we obtain:

$$\mu_{ip} = \frac{e^{\lambda_i[\eta_p - (-\nu_i/\lambda_i)]}}{1 + e^{\lambda_i[\eta_p - (-\nu_i/\lambda_i)]}} \tag{11}$$

Equation 11 can be seen as a reparameterization of the 2PLM (see Equation 3) where $\nu_i = -\alpha_i \delta_i$ and $\lambda_i = \alpha_i$.

The MNLFA extends the GLFA model by also allowing model parameters to vary across different levels of the included covariates. For the sake of simplicity and without loss of generality, I consider a unidimensional model with one continuous background variable x_i. Within the MNLFA framework, the test for measurement invariance is done by including the background variable as a moderator in the model (Bauer, 2017), see figure 2. Here it is shown that the background variable x can moderate any of the model parameters of the factor model given in the dashed rectangle. If it has an effect on the item parameters, measurement invariance does not hold (Bauer, 2017). For each parameter in the model, a moderation function has to be specified. Following Bauer and Hussong (2009) linear functions for the intercepts and factor loadings are defined.

The intercept ν_{ip} for item i and person p is defined as follows:

$$\nu_{ip} = \nu_{0i} + \kappa_i x_p, \tag{12}$$

where ν_{0i} is the baseline intercept (when the covariate is zero) for item i and κ_i represents the effect of the covariate x_p on the intercept. Similarly, the factor loading λ_{ip} for item i and person p is defined as follows:

$$\lambda_{ip} = \lambda_{0i} + \omega_i x_p, \tag{13}$$

where λ_{0i} is the baseline factor loading for item i and ω_i represents the effect of the covariate x_p on the factor loading. For the mean and variance of the latent factor η_p, moderation functions are defined as well. However, if these moderations are significant this is not an indication of a violation of measurement invariance. Rather, this means the background variable has some effect on the latent construct, which is referred to as *impact* (Bauer et al., 2020). For the factor mean a linear moderation can be defined as follows:

$$\alpha_p = \alpha_0 + \gamma x_p, \tag{14}$$

where α_p is the factor mean and γ represents the effect of the covariate x_p on the factor mean. For the factor variance we cannot define a linear function, since this could violate the requirement that variances must be positive. Instead, a log-linear moderation is used (following Bauer (2017) and Bauer et al. (2020)):

$$\phi_p = \phi_0 + e^{\beta x_p}, \tag{15}$$

where ϕ_p is the factor variance and γ represents the effect of the covariate x_p on the factor mean. To include the specified moderations in Equation 11, the correct subscripts are added to the parameters:

$$\mu_{ip} = \frac{e^{\lambda_{ip}[\eta_p - (-\nu_{ip}/\lambda_{ip})]}}{1 + e^{\lambda_{ip}[\eta_p - (-\nu_{ip}/\lambda_{ip})]}} \tag{16}$$

In the same way as before, Equation 16 can be seen as a reparameterization of the 2PLM. If we consider the 1PLM, as was done by Bechger and Maris (2014), we would simply set the factor loadings/discrimination parameters to be 1 for all items. Equation 16 then simplifies to:

$$\mu_{ip} = \frac{e^{\eta_p + \nu_{ip}}}{1 + e^{\eta_p + \nu_{ip}}} \tag{17}$$

Equation 17 can now be seen as a reparameterization of the 1PLM (Equation 2) where $\nu_{ip} = -\delta_{ip}$, we will refer to this model as the MNLFA-1PLM. From Equation 12 it follows that:

$$\delta_{ip} = -\nu_{0i} - \kappa_i x_p \tag{18}$$

Thus, for all i items we obtain a baseline item difficulty/intercept ν_{0i} and an effect of the covariate x_p given by κ_i.

2.3 Relative difficulties in the MNLFA-1PLM

Since we are considering *relative* difficulties, the interest might not appear to be in whether or not $\kappa_i = 0$, for all i. Rather, the focus would be on testing if the effect of the grouping variable x_j is consistent across all items. Thus, if $\kappa_1 = \kappa_2 = ... = \kappa_i$. However, we cannot arrive at a situation where $\kappa_1 = \kappa_2 = ... = \kappa_i = c$, where c is some constant. Since this constant effect c of x_j should be apparent not on the item difficulties but on the factor means (see γ in Equation 14). So, for our overall test of DIPF, the interest is actually in testing $\kappa_i = 0$, for all i. Thus, for this overall test, existing MNLFA testing procedures can be used (see Kolbe, 2022).

However, once DIPF is detected in the test, this perspective changes. After rejecting the hypothesis of no-DIPF across a test, the next step would be to see if κ stays consistent across pairs of items. Instead of looking at item-pairs, we follow the concept of item clusters as defined by Bechger and Maris (2014). A cluster is a set of items for which the relative difficulties, within that cluster, are invariant. Testing whether clusters are invariant can provide an alternative to testing every possible item-pair in a test. So, the interest lies in testing if the effect of the grouping variable x_p is consistent across a specific cluster of items. Thus, for items $1, 2, ..., m$ within a cluster we want to know if $\kappa_1 = \kappa_2 = ... = \kappa_m$. Bechger and Maris (2014) suggest the use of heatmaps showing all item-pair comparisons $\hat{D}_{ij}$ (see Equation 6) to identify possible clusters to use in testing for invariant clusters, when no a priori hypothesis about possible item clusters exist.

Condition	n	DIF strength	# DIF items
1	500	small	1
2	500	small	5
3	500	small	7
4	500	large	1
5	500	large	5
6	500	large	7
7	1,000	small	1
8	1,000	small	5
9	1,000	small	7
10	1,000	large	1
11	1,000	large	5
12	1,000	large	7
13	5,000	small	1
14	5,000	small	5
15	5,000	small	7
16	5,000	large	1
17	5,000	large	5
18	5,000	large	7

Table 1: Simulation conditions. Strength of DIF was varied from small to large by using weighted area between the curves.

3 Method

A test for testing DIPF across item clusters using a continuous background variable in the MNLFA model will be developed. Then, code will be written to implement this test using R (version 4.2.1; R Core Team, 2022). This will done by using the package *mnlfa* (version 0.2.4; Robitzsch, 2022), which allows for the implementation of the MNLFA-1PLM in R. Thus, estimates of the κ_i's can be obtained using this package.

A simulation study will be conducted to assess the power of the DIPF-MNLFA test, i.e. its ability to correctly identify DIPF. The data generating mechanism employed mimics that off the Bauer et al. (2020) MNLFA study. First, a continuous covariate was sampled from a normal distribution with $X \sim N(25, 5)$. This could be interpreted as representing age, for example. Second, latent scores are generated using Equations 14 and 15. Third, individual item scores were generated under the MNLFA-1PLM (see Equation 17 and 12). Parameters used in these last two steps will differ across the different simulation conditions (see below).

We keep the number of items and the number of invariant clusters constant (15 items and two invariant clusters). Both of these choices are essentially arbitrary, they were made for simplicity and to keep the number of simulation conditions limited. The choice for 15 number of items was based on other simulation studies (namely, Bauer et al., 2020; Bechger and Maris, 2014; Kim et al., 2017) where the number of items in a test ranged from 6 to 30. The focus will be on the effects of: sample size (500 - 5,000), strength of DIF (small/large), and number of DIF items (i.e. the size of the smaller cluster, 1 - 7). Specifications of the simulation conditions are provided in table 1. Strength of DIF was varied from small to large by using weighted area between the curves, again following Bauer et al. (2020).

References

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bauer, D. (2017). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

Bauer, D., 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), 43–55. https://doi.org/10.1080/10705511.2019.1642754 _eprint: https://doi.org/10.1080/10705511.2019.1642754

Bauer, D., & Hussong, A. (2009). Psychometric Approaches for Developing Commensurate Measures Across Independent Studies: Traditional and New Models. *Psychological methods*, *14*, 101–25. https://doi.org/10.1037/a0015583

Bechger, T., & Maris, G. (2014). A Statistical Test for Differential Item Pair Functioning. *Psychometrika*. https://doi.org/10.1007/s11336-014-9408-y?sa_campaign=email/event/articleAuthor/onlineFirst#

Cuellar Caicedo, E. (2022). *Making sense of DIF in international large-scale assessments in education* (Doctoral dissertation).

Glas, C., & Ouborg, M. (1993). Vraagonzuiverheid. *Psychometrie in de Praktijk*, 349–371.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Janssen, R. (2018). Using a Differential Item Functioning Approach to Investigate Measurement Invariance. In *Cross-Cultural Analysis* (Second). Routledge.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 22.

Kolbe, L. (2022). *Novel Approaches to Assess Measurement Invariance* (Doctoral dissertation).

Kush, J. M., Masyn, K. E., Amin-Esmaeili, M., Susukida, R., Wilcox, H. C., & Musci, R. J. (2022). Utilizing Moderated Non-linear Factor Analysis Models for Integrative Data Analysis: A Tutorial. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–16. https://doi.org/10.1080/10705511.2022.2070753 _eprint: https://doi.org/10.1080/10705511.2022.2070753

Lord, F. M. (1976). *A Study of Item Bias Using Characteristic Curve Theory.*

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. Routledge. https://doi.org/10.4324/9780203821961

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Robitzsch, A. (2022). *Mnlfa: Moderated nonlinear factor analysis* [R package version 0.2-4]. https://CRAN.R-project.org/package=mnlfa

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Weston, R., & Gore, P. (2006). A Brief Guide to Structural Equation Modeling. *The Counseling Psychologist*, *34*, 719–751. https://doi.org/10.1177/0011000006286345