MIMIC models for uniform and non-uniform DIF as moderated mediation models

—

THIS MANUSCRIPT IS IN PRESS AT *APPLIED PSYCHOLOGICAL METHODS*

DOI: 10.1177/0146621619835496

Amanda K. Montoya
University of California, Los Angeles

Minjeong Jeon
University of California, Los Angeles

Author Note

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE-1343012. Correspondence concerning this article should be addressed to Amanda K. Montoya, Department of Psychology, University of California - Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095. Email: akmontoya@ucla.edu.

Abstract

In this note we describe how multiple indicators multiple cause (MIMIC) models for studying uniform and non-uniform differential item functioning (DIF) can be conceptualized as mediation and moderated mediation models. Conceptualizing DIF within the context of a moderated mediation model helps us understand DIF as the effect of some variable on our measurements which is not accounted for by the latent variable of interest. In addition, this allows us to apply useful concepts and ideas from the mediation and moderation literature: (1) improving our understanding of uniform and non-uniform DIF as direct effects and interactions, (2) understanding the implication of indirect effects in DIF analysis, (3) clarifying the interpretation of the "uniform DIF parameter" in the presence of non-uniform DIF, and (4) probing interactions and using the concept of "conditional effects" to better understand the patterns of DIF across the range of the latent variable.

Keywords: Differential item functioning, item response theory, MIMIC models, mediation, moderated mediation

MIMIC models for uniform and non-uniform DIF as moderated mediation models

—

THIS MANUSCRIPT IS IN PRESS AT *APPLIED PSYCHOLOGICAL METHODS*

DOI: 10.1177/0146621619835496

1 Introduction

One of the primary aims of measurement research is to develop and identify valid sets of items which measure a specific latent variable. Much research in measurement, particularly within item response theory (IRT), focuses on differential item functioning (DIF) as it can be a major threat to the validity of a scale or set of items. DIF occurs when the probability of a response on a specific item is dependent on some external factor even after conditioning on the latent trait. For instance, it is crucial in educational assessment that a mathematics test is equally valid for all students taking the test. This means that controlling for a student's math ability, there should be no other factors which increase or decrease the probability of getting a item correct.

Zumbo (2007) deemed the "Third Generation" of DIF research to be an era where we investigate and aim to understand why DIF is occurring rather than just detect or correcting for it. One of the primary aims of this note is provide a more intuitive understanding of uniform and non-uniform DIF using a MIMIC model and the concepts of indirect effects, direct effects, and interactions commonly used in mediation and moderation. We believe that this type of framework may help researchers to understand DIF in a way that would facilitate thinking about why or how DIF occurs, by focusing on DIF as a process rather than a nuisance.

1.1 Uniform and Non-Uniform DIF

Differential item functioning can occur in two primary ways: uniform and non-uniform DIF (Mellenbergh, 1982). Figure 1 gives four examples of differential item functioning. Early definitions of uniform and non-uniform DIF emphasized the mutually exclusive nature of these two types of differential item functioning (Mellenbergh, 1982; Ackerman, 1992; Millsap & Everson, 1993; Narayanan & Swaminathan, 1996). For example Ackerman (1992) defined uniform DIF as when the "ICCs for the different groups differ by only a horizontal translation (i.e., they are parallel but not coincident)." Alternatively, non-uniform DIF is when the ICCs are nonparallel. Using these definitions of uniform and non-uniform DIF Panel A in Figure 1 is the only item which would fit the definition of uniform DIF, all other panels would have non-uniform DIF, because the ICCs are not parallel. We rely on definitions of uniform and non-uniform DIF such that they are mutually exclusive. Additionally we describe how to interpret the parameter which in models without non-uniform DIF is used to describe uniform DIF, in the presence of non-uniform DIF. This parameter is what differentiates Panels B, C, and D in Figure 1.

[Figure 1 about here]

1.2 MIMIC Models for DIF Testing

A number of methods have been developed for detecting and investigating DIF including Mantel-Haenszel tests (Holland & Thayer, 1988), multidimensional approaches (e.g., SIBTESTs; Shealy & Stout, 1993), area methods (Raju, 1988, 1990), logistic regression (Swaminathan & Rogers, 1990), multiple group IRT and structural equation models (Jöreskog, 1971), and multiple indicators multiple cause (MIMIC) models (B. O. Muthèn, 1985, 1989; B. O. Muthèn, Kao, & Burstein, 1991). For reviews of these methods see Magis, Bèland, Tuerlinckx, and De Boeck (2010) and Zumbo (2007). In this article, we will focus on the method of using MIMIC models for studying DIF. MIMIC models are characterized by the inclusion of independent variables (or causes, $\mathbf{X}$) of a latent variable and its observed indicators ($\mathbf{Y}$) (Hauser & Goldberger, 1971; Joreskog & Goldberger, 1975). Figure 2 displays an example MIMIC model with a case of three observed indicators and three observed causes.

[Figure 2 about here]

MIMIC models were first used to study DIF by B. O. Muthèn (1985), who showed that a typical two-parameter IRT model assumes no direct path between an independent variable (cause) and an individual item response (indicator). A path between an independent variable and an item can be included in a MIMIC model. If this direct path is found to be statistically significant, this suggests that for two people matched on the latent variable, the item is easier for one of those people. This is an indication of uniform DIF.

The accuracy of MIMIC models for detecting DIF has been investigated by Woods (2009) and Woods and Grimm (2011) and others. Researchers have also compared MIMIC methods to other DIF detection methods such as the Mantel-Haenszel test, SIBTEST, and likelihood-ratio tests based on multiple group IRT models (Finch, 2005; Woods, 2009; Woods & Grimm, 2011). Simulation studies have found that the MIMIC method works similarly to other tests, but requires smaller sample sizes than other methods to detect uniform DIF (Woods, 2009).

The MIMIC approach is not immune to many of the issues common to most DIF testing techniques. Scale purification is a procedure for detecting which items among a set of items have DIF (Lord, 1980). This process is incredibly important, as in scale development researchers often have a set of items and they want to detect which items within the set have DIF, rather focusing on a single item. Wang, Shih, and Yang (2009) studied the performance of an iterative purification method with MIMIC models and found this method worked well for identifying items with DIF. However like many other scale purification methods, when there are too many items with DIF the scale purification methods have Type I error rates which are too high to be acceptable (Wang et al., 2009; Shih & Wang, 2009). Shih and Wang (2009) proposed a method for identifying a short anchor (a small set of DIF free items). This method seems to perform better than iterative purification methods when the proportion of items with DIF is large. Indeed only one DIF free item was needed to control Type I Error; however, longer anchors increased power to detect DIF. Additionally Shih and Wang (2009) showed how a slight alteration of the iterative methods proposed by Wang et al. (2009) could be used to identify anchor items (rather than identifying DIF items). In comparison to other DIF detection methods, MIMIC methods are less susceptible to DIF in anchor items (Finch, 2005).

Both MIMIC models and multidimentional IRT models can be used to relax some of the more stringent assumptions of traditional IRT models, such as unidimensionality (Wang et al., 2009; Zumbo, 2007; Lee, Bulut, & Suh, 2016). In a MIMIC model, by allowing individual items to load on multiple latent variables, we can allow for multidimensionality while testing for DIF. There are many instances where researchers attribute DIF to multidimensionality. Cheng, Shao, and Lathrop (2016) proposed a method for understanding DIF using multidimensional MIMIC models. If there are indicators of a latent construct which the researchers believes to be causing DIF, a MIMIC model can be estimated including that latent construct and testing whether there is DIF above and beyond this additional latent construct. Alternatively, if there is no remaining DIF this would support the claim that the additional latent construct explains all (or most of) the DIF (Cheng et al., 2016). The Mantel-Haenszel test and SIBTEST are not appropriate for multidimensional modeling because they are designed for only one latent trait (Bulut & Suh, 2017); however, a multidimensional SIBTEST has been developed (Stout, Li, Nandakumar, & Bolt, 1997).

MIMIC models are useful in that they can be used for continuous, categorical, or mixes of continuous and categorical outcomes. Throughout this manuscript we provide examples and equations for dichotomous outcomes only. Because MIMIC models are estimated in a structural equation modeling framework model fit indices are available for these models. However, many model fit statistics are only valid for models with continuous outcomes (e.g., RMSEA, SRMR, CFI, TLI; Yun, 2002). Other measures such as information criteria can be used to compare models with dichotomous outcomes (Kang & Cohen, 2007).

1.3 Aims and Structure

In DIF studies researchers are often concerned about what characteristics of an individual (e.g., gender, race, socioeconomic status) might lead to uniform or non-uniform DIF. However, applied researchers seem to pay little attention to models of how independent variables affect measurement and do not include DIF as part of their model. By viewing MIMIC models as mediation models, it becomes clear that there are multiple ways an independent variable can affect an item response, not all of which undesirable.

The purpose of this note is to show that MIMIC models for uniform and non-uniform DIF analysis can be conceptualized as mediation and moderated mediation models. Mediation and moderation models provide a framework for understanding the process through which some effect occurs and for modeling contingencies in those processes. An advantage of understanding DIF in the context of mediation and moderation analysis is that it provides us with an opportunity to investigate the mechanisms through which independent variables influence our measurements. What we find is that DIF is one of these processes. In addition, conceptualizing within a mediation and moderation framework allows us to apply useful concepts and ideas from the mediation and moderation literature to improve our understanding of DIF. This includes (1) to appreciate uniform and non-uniform DIF as direct effects and interactions, (2) to understand the implication of indirect effects in DIF analysis, (3) to revise a conventional interpretation of the uniform DIF parameter in the presence of non-uniform DIF, and (4) use the concept of probing interactions to better understand the patterns of DIF.

The remainder of this note is structured as follows: In Section 2, we will first show how MIMIC models for uniform and non-uniform DIF analysis can be understood within a mediation and moderation framework. In the subsequent section (Section 3), we will discuss in detail how some important concepts and ideas from the mediation and moderation literature can be applied to improve our understanding of DIF. Throughout Sections 2 and 3 we describe the analysis of a dataset to show how MIMIC DIF analysis can be done and to provide a concrete example for the advantages of using mediation and moderation ideas to interpret DIF. We will then end in Section 4 with some concluding remarks.

2 MIMIC DIF Models as Mediation and Moderated Mediation Models

To illustrate how MIMIC DIF models can be understood within a mediation and moderation framework, we use a dataset which explored cohort differences in intelligence testing on samples of children, age 12 to 14, from Estonia (Must & Must, 2014)[1]. The first cohort was collected during 1933 - 1936 ($N = 890$) and the second cohort was collected in 2006 ($N = 913$). The study focused on differences in intelligence across cohorts, as measured by ten subtests which cross a variety of domains. For simplicity we will only examine one subtest: arithmetic, with 16 items. Each student responded with an open response, these responses were given a binary coding: 1 (correct) or 0 (incorrect). The 16 items were administered in Estonian (Haggerty, Terman, Thorndike, Whipple, & Yerkes, 1921); Approximate English translations for each item are in Table 1 along with proportions of correct answers for each cohort.

For 10 of 16 of the items, the 2006 cohort performed better than the 1933/36 cohort. The exceptions where the 1933/36 cohort outperformed the 2006 cohort include Items 5, 9, 10, 11, 12, and 13. We will use this data to explore cohort differences in latent arithmetic ability as well as the potential for differential item functioning. There are a variety of counter-explanations for why the later cohort outperformed the earlier cohort, and these are discussed in Must and Must (2013). These analyses are not to be taken as novel theoretical findings, but rather an example used for showing how to conduct DIF analysis. We selected two items to demonstrate how to test for DIF with MIMIC models: Items 5 and 10. These items were selected because they show interesting patterns of DIF as will be seen in later sections. The focus of this manuscript is on the estimation of DIF within a single item; however when these analyses are done in empirical data we recommend that when doing this type of analysis researchers should use the methods proposed by Shih and Wang (2009) or Wang et al. (2009) as described in Section 1.2. Alternatively researchers can use substantive knowledge to inform which items to explore for DIF. We did not conduct any scale purification for these analyses. Data analysis was done using Mplus Version 8.1 (L. K. Muthèn & Muthèn, 1998 – 2011). For this analysis we used maximum likelihood estimation with robust standard error estimates. The model was unidimensional, assuming that all items loaded onto a single latent variable which we call arithmetic ability. Select input files are included in the appendices to aid implementation.

[Table 1 about here]

2.1 Mediation Model For Uniform DIF

In the example we are concerned with whether there are cohort differences on the probability of correctly answering an item. One reason that there may be differences in the

[1] This data is freely available for download through the Journal of Open Psychology Data at http://doi.org/10.7910/DVN/23791

probability of responding is differences in ‘true’ latent ability. However, there may be concern that for a specific item there are cohort differences in the probability of responding that are not attributable to ‘true’ differences in latent ability. For this reason we add a direct path between the cohort variable and the item response, as represented in Figure 3a.

[Figure 3 about here]

For notation, we use Y_{it} to denote a binary response to item i ($i = 1, ..., I$) for person t ($t = 1, ..., N$), X_t to indicate the independent variable that indicates 1933/36 cohort (coded as 0) and 2006 cohort (coded as 1), and θ_t to represent a latent trait (‘true’ arithemtic ability) that we want to measure using the I items (in the example $I = 16$). This MIMIC model with a direct path can be viewed as a mediation model that allows the effect of X_t on $P(Y_{it})$ [2] to be mediated by θ_t. We will specify two regression models that include the three paths in Figure 3a (paths [a], [b], and [c]).

First, the model for path [a] is specified with the latent trait θ_t as the dependent variable:

$$\theta_t = \delta X_t + \xi_t, \tag{1}$$

where the regression coefficient δ represents the influence of the variable X_t on the θ_t and we assume the residual ξ_t follows a normal distribution with $\xi_t \sim N(0, \sigma^2)$. In the arithmetic example, the δ parameter indicates the mean difference between the 1933/36 cohort and the 2006 cohort on ‘true’ arithmetic ability (θ_t), which is called *impact* in the measurement literature (Ackerman, 1992; Camilli, 1993).

Estimating Equation 1 with the arithmetic data and allowing Item 5 to have uniform DIF provides an estimate of δ, 0.149. This means that the 2006 cohort is on average 0.149 units higher than the 1933/36 cohort on latent arithmetic ability ($p = 0.051$). This effect can be interpreted in standard deviation units, since the latent variable is standardized with a variance of 1. The effect is relatively small and only nears statistical significance when using a relatively generous level of 0.05.

Second, a model for path [b] and [c]:

$$P(Y_{it}|\theta_t, X_t) = g^{-1}(\beta_i + \gamma_i^{\beta} X_t + \alpha_i \theta_t), \tag{2}$$

where g^{-1} indicates the inverse of a logit link function[3]. The coefficient α_i denotes the

[2] For notational simplicity we denote the probability that the outcome Y for item i and individual t is 1 as $P(Y_{iy})$

[3] The link function can be either logit or probit for binary item response data. We will use a logit link function for consistent discussions throughout the paper.

influence of the latent variable θ_t on $P(Y_{it})$, controlling for X_t. The coefficient γ_i^β denotes the influence of the independent variable X_t on $P(Y_{it})$ while controlling for the latent variable. In the mediation literature, this is called the *direct effect*, because it is the effect of the variable X_t on the outcome, not through the mediator. Here the mediator θ_t is a continuous latent variable with a distributional assumption, $\theta_t \sim N(0, 1)$. Note that Equation 2 corresponds to the MIMIC model for uniform DIF (Woods, 2009) where γ_i^β corresponds to the uniform DIF parameter for item i.

Estimating Equation 2 for Item 5 we get:

$$P(Y_{5t}|\theta_t, X_t) = g^{-1}(1.28 - 0.68X_t + 1.03\theta_t),$$

The influence of θ_t on the probability of correct response on Item 5 is positive ($\hat{\alpha}_5 = 1.03$, $SE = 0.121$, $p < .001$). This means that those who have a higher latent ability are more likely to get this item correct. However, the estimate of the effect of group on Item 5 is negative, ($\hat{\gamma}_5^\beta = -0.68$, $SE = 0.15$, $p < .001$). This suggests controlling for latent ability, individuals in the 2006 cohort are less likely to answer this item correctly than individuals in the 1933/36 cohort. This is uniform differential item functioning or, in the mediation framework, a direct effect. Item 5 was one of the items where the 2006 cohort performed worse than the 1933/36 cohort. An approximate translation of the item is "How much longer is 12 yards than a meter?" It is difficult to stipulate why Item 5 performs this way. Must and Must (2014) noted that the 1930s cohort has more students who were rural than the 2006 cohort. Perhaps rural children were more familiar with distances like yards and meters.

A similar analysis can be conducted on Item 10 which is another item where the 1933/36 cohort performed better than the 2006 cohort. The results showed that $\hat{\delta} = 0.09$, $SE = 0.079$, $p = 0.26$ [4]. In this analysis there is less convincing evidence that the two cohorts significantly differ on latent arithmetic ability. Based on the estimated model for Equation 2, the estimated effect of θ_t on the probability of a correct response on Item 10 is 0.94, $SE = 0.13$, $p < 0.001$. Individuals with higher latent arithmetic ability are more likely to answer Item 10 correctly. Controlling for latent ability, there is not a significant difference between cohorts on the probability of answering Item 10 correctly $\hat{\gamma}_{10}^\beta = -0.16$, $SE = 0.15$, $p = .29$. There is not sufficient evidence of a direct effect (i.e., uniform DIF) for Item 10.

By plugging in Equation 1 as θ_t in Equation 2 we can get additional information about indirect and direct effects .

[4] Because Equations 1 and 2 are estimated simultaneously, the estimates of δ differ depending on which items are allowed to have DIF. It is also possible to allow both items to have DIF, but for simplicity we do not explore that option.

$$\begin{aligned} P(Y_{it}|\theta_t, X_t) &= g^{-1}(\beta_i + \gamma_i^{\beta} X_t + \alpha_i(\delta X_t + \xi_t)), \\ &= g^{-1}(\beta_i + \gamma_i^{\beta} X_t + \alpha_i(\delta X_t + \xi_t)), \\ &= g^{-1}(\alpha_i \delta X_t + \gamma_i^{\beta} X_t + \alpha_i \xi_t + \beta_i), \\ &= g^{-1}((\underbrace{\delta\alpha_i}_{\text{indirect effect}} + \underbrace{\gamma_i^{\beta}}_{\text{direct effect}})X_t + \alpha_i \xi_t + \beta_i). \end{aligned} \tag{3}$$

Equation 3 shows how the effect of X_t on $P(Y_{it})$ is parsed into two effects: a direct effect (γ_i^{β}) and an indirect effect ($\delta\alpha_i$). Specifically, the indirect effect quantifies the effect of X_t on $P(Y_{it})$ through the mediator θ_t (represented by paths [a] and [b]), while the direct effect quantifies the remaining effect of X_t on $P(Y_{it})$ (through path [c]). From the above exercise, we can see that specifying a MIMIC model with a direct path (for uniform DIF) can be viewed as as a mediation model, while this perspective provides an opportunity to discuss two potential mechanisms (direct and indirect routes) that generates group differences on $P(Y_{it})$.

Our uniform DIF examples for Item 5 and Item 10 can be used to generate estimates of indirect effects. The estimated indirect effect of cohort on Item 5 through latent arithmetic ability is $0.149 \times 1.03 = 0.153$. The estimated indirect effect of cohort on Item 10 through latent arithmetic ability is $0.09 \times 0.94 = 0.084$. We discuss how these estimates can be interpreted in Section 3.2.

2.2 Moderated Mediation Model For Non-Uniform DIF

It may be possible that some items on the questionnaire provide more information about the latent abilities of the individuals in one of the cohorts compared to the other. By that we mean that the items' ability to discriminate among people of different latent abilities may depend on which cohort those people come from. For instance, as a latent ability increases, a child's probability of correctly answering the item might increase faster if that child is in the 1933/36 cohort compared to the 2006 cohort. This type of effect is described as non-uniform DIF. To test this hypothesis, we allow the path between θ_t and $P(Y_{it})$, as specified in Section 2.1, to be a function of X_t in the MIMIC model. This modification is displayed in Figure 3b.

This revised MIMIC model is also a moderated mediation model. Modifying the regression model for paths [b] and [c] by including the interaction between X_t and θ_t, provides an additional parameter represented by path [d] as follows:

$$\begin{aligned} P(Y_{it}|\theta_t, X_t) &= g^{-1}(\beta_i + \gamma_i^{\beta} X_t + \alpha_i \theta_t + \gamma_i^{\alpha} X_t \theta_t), && (4) \\ &= g^{-1}(\beta_i + \gamma_i^{\beta} X_t + (\alpha_i + \gamma_i^{\alpha} X_t)\theta_t). && (5) \end{aligned}$$

Note that Equation 4 includes the interaction between X_t and θ_t. Equation 5 shows how the relationship between θ_t and $P(Y_{it})$ now depends on X_t. If the parameter γ_i^{α} is zero, then the relationship between θ_t and $P(Y_{it})$ does not depend on X_t. Hence, the coefficient γ_i^{α} indicates the moderation of the effect of θ_t on $P(Y_{it})$ by X_t. Note that Equation 4 corresponds to the MIMIC model for studying non-uniform DIF (Woods & Grimm, 2011), where γ_i^{α} corresponds to the non-uniform DIF parameter for item i and $\theta_t = \delta X_t + \xi_t$ with $\xi_t \sim N(0, \sigma^2)$. In the arithmetic example, γ_i^{α} can be interpreted as a difference in the relationship between 'true' arithmetic ability and the probability of answering correctly on item i for the 2006 cohort compared to the 1933/36 cohort. So if this parameter is positive the item is more discriminating for the 2006 cohort. If it is negative it is more discriminating for the 1933/36 cohort. Similarly we can think estimating Equation 2 just the 1933/36 cohort and again with just the 2006 cohort, γ_i^{α} would indicate the difference in the α_i parameters between the two cohorts.

Estimating Equation 1 and 4 for Item 5 we get an estimate of δ (0.147) and an estimated equation:

$$P(Y_{5t}|\theta_t, X_t) = g^{-1}(1.20 - 0.57X_t + 0.75\theta_t + 0.68X_t\theta_t).$$

As will be discussed more in depth in Section 3.4 the coefficient for θ_t is no longer the overall effect of the latent variable on the probability of correct response. Now it is the effect of the latent variable on the probablly of responding correctly when $X_t = 0$ (i.e., for individuals in the 1933/36 cohort). So for individuals in the 1933/36 cohort, θ_t positively predicts the probability of a correct response on Item 5 ($\hat{\alpha}_5 = 0.75$, $SE = 0.14$, $p < 0.001$). As mentioned previously the γ_5^{α} parameter denotes the differences in the item discrimination parameter across the cohorts. Alternatively, we can think of this parameter as the degree to which X_t moderates the relationship between the latent variable, θ_t and the probability of correct response, $P(Y_{it})$. So for individuals in the 2006 cohort the item discrimination parameter is $0.75 + 0.68 = 1.43$. The test on the difference suggests that Item 5 has a greater item discrimination parameter for the 2006 cohort compared to the 1933/1936 cohort ($\hat{\gamma}_5^{\alpha} = 0.68$, $SE = 0.27$ $p = .012$). The coefficient for X_t represents the effect of cohort on the probability of correct response for individuals average on the latent arithmetic trait. This suggests for individuals who are average on the latent arithmetic trait, those in the 1933/36 cohort are more likely to answer the item correctly ($\hat{\gamma}_5^\beta = -0.57$, $SE = 0.16$, $p < .001$).

We conducted the same analysis, allowing for non-uniform DIF for Item 10 only. The estimated cohort difference on the latent trait is 0.09 ($SE = 0.08$, $p = .270$). The effect of the latent trait on the probability of answering Item 10 correctly for individuals in the 1933/36 cohort is 1.601 ($SE = .30$, $p < 0.001$). So for this cohort increased latent ability predicts a higher probability of correctly answering the item. But for the 2006 cohort the item discrimination parameter is much lower: $1.601 - 0.987 = 0.614$. The delta method can be used to test if the item discrimination parameter is significantly different from zero[5]. The results show that the item discrimination parameter for the 2006 cohort is still significantly different from zero ($SE = 0.146$, $p < .001$). The test on $\hat{\gamma}_{10}^\alpha$ suggests that the item discrimination parameter is significantly lower for the 2006 cohort compared to the 1933/36 cohort ($SE = 0.335$, $p = 0.003$). This suggests that the strength of the relationship between latent ability and probability of correct response is moderated by cohort. For individuals at the mean of the latent arithmetic trait, there was no sufficient evidence that there were cohort differences in the probability of answering Item 10 correctly ($\hat{\gamma}_i^\beta = -0.036$, $SE = 0.163$, $p = 0.827$).

Just as we did with the models for uniform DIF we can combine Equation 5 with Equation 1 to get information about indirect and direct effects,

$$\begin{aligned}
P(Y_{it}|\theta_t, X_t) &= g^{-1}((\alpha_i + \gamma_i^\alpha X_t)\theta_t + \gamma_i^\beta X_t + \beta_i), \qquad (6)\\
&= g^{-1}((\alpha_i + \gamma_i^\alpha X_t)(\delta X_t + \xi_t) + \gamma_i^\beta X_t + \beta_i)),\\
&= g^{-1}((\underbrace{(\delta\alpha_i + \delta\gamma_i^\alpha X_t)}_{=\underbrace{(\alpha_i + \gamma_i^\alpha X_t)\delta}_{\text{indirect effect}}} + \underbrace{(\gamma_i^\alpha \xi_t + \gamma_i^\beta)}_{\text{direct effect}})X_t + \alpha_i\xi_t + \beta_i)), \qquad (7)
\end{aligned}$$

Equation 7 is equivalent to a moderated mediation model that shows that the indirect effect of X_t on $P(Y_{it})$ through θ_t is a function of X_t. To understand this, see the first term in Equation 7, which represents the indirect effect, can be re-written as $(\delta\alpha_i + \delta\gamma_i^\alpha X_t) = (\alpha_i + \gamma_i^\alpha X_t)\delta$. The fact that the indirect effect depends on a categorical grouping variable X_t means that the indirect effect may differ across groups. Note that if $\delta\gamma_i^\alpha = 0$ in Equation 7, the indirect effect no longer depends on X_t, meaning that the indirect effect is the same across groups. This means that if $\delta\gamma_i^\alpha \neq 0$ there is no single indirect effect

[5] Including a new parameter in the "Model Constraint" section of Mplus code will do this automatically. See Appendices.

that applies to all groups of people and group-specific indirect effects should be estimated.

For instance, in the arithmetic example with Item 5, the indirect effect is $\delta\alpha_i = 1.47 \times 0.75 = 1.102$ for the 1933/36 cohort ($X_t = 0$) and $\delta\alpha_i + \delta\gamma_i^\alpha = 1.47 \times .75 + 1.47 \times 0.68 = 2.102$ for the 2006 cohort ($X_t = 1$). In the mediation literature $\delta\gamma_i^\alpha$ is called the index of moderated mediation which quantifies the change in the indirect effect with a one unit change in the moderator, in this case X_t (Hayes, 2015).

3 Applying Ideas of Moderation and Mediation to DIF

In Section 2, we discussed that MIMIC models for studying uniform or non-uniform DIF can be viewed as mediation or moderated mediation models, respectively. Specifically, the uniform DIF parameter γ_i^β from Equation 2 corresponds to a direct effect in a mediation model. This means that uniform DIF indicates the degree to which the independent variable X_t has an effect on the outcome $P(Y_{it})$ not through the mediator θ_t. In addition, the non-uniform DIF parameter γ_i^α from Equation 6 corresponds to an interaction effect. This implies that non-uniform DIF indicates the degree to which the effect of the mediator θ_t on the outcome variable $P(Y_{it})$ is moderated by the independent variable X_t.

In this section, we will discuss how some of additional ideas from the mediation and moderation literature can improve our understanding of DIF.

3.1 Null Direct Effects

In mediation analysis, a null direct effect implies that the cumulative effects of the other ways (not through the mediator) in which an independent variable (X) affects the outcome (Y) average out to be zero. It does not mean that the independent variable has no impact on the outcome through any other factors (Rucker, Preacher, & Tormala, 2011; Hayes, 2018). Suppose the effect of an independent variable on an outcome is mediated through three variables ($\mathbf{M}$) where each indirect effect amounts to 2, -2, and 2, respectively. If a researcher specifies a model with only one mediator (with an indirect effect of 2), the effects of the other two mediators (-2 and 2) will average out to be zero. Accordingly, the direct effect of X on Y is zero in this case, even though there are additional ways that X affects Y.

Similarly, a zero uniform DIF parameter value ($\gamma_i^\beta = 0$) implies that the effects of all other factors which may mediate the effect of X_t on $P(Y_{it})$ are canceled out. For instance, in the arithmetic example with Item 10, we originally found that the uniform DIF parameter was not significantly different from zero. Suppose the we considered farming experience as an additional mediator in studying uniform DIF for Item 10. With such an inclusion of a relevant mediator, the direct effect of X_t on $P(Y_{it})$ (or uniform DIF) may no longer be zero. This means that Item 10 may be equally difficult across the cohorts among those matched on ‘true’ arithmetic ability. But the item may not be equally difficult among individuals in the two cohorts matched on ‘true’ arithmetic ability *as well as* farming experience.Just as mediation experts do not recommend interpreting a null direct effect as implying that all mediators have been accounted for, similar cautions should be taken when interpreting a lack of uniform DIF.

3.2 Indirect Effects

An indirect effect in Figure 3a (paths [a] and [b]) represents the effect of the independent variable (X_t) on the outcome ($P(Y_{it})$) through the mediator (θ_t). That is, the indirect effect measures the degree to which a group difference on the latent trait θ_t contributes to a group difference on the probability of endorsing item i. The indirect effect can be computed as the product of the two coefficients for paths [a] and [b], that is, $\delta\alpha_i$.

We discussed in Section 2.1 that the effect of X_t on $P(Y_{it})$ is additively decomposed into a direct effect (represented by the uniform DIF parameter γ_i^β) and an indirect effect (represented by the product of the two parameters $\delta\alpha_i$). This means that the indirect effect ($\delta\alpha_i$) carries information about group differences on $P(Y_{it})$ which can be explained only by group differences in θ_t and the relationship between θ_t and $P(Y_{it})$.

Both in the presence and absence of a direct effect (or uniform DIF), it may be worthwhile to examine and understand the indirect effect. Consider the uniform DIF MIMIC model. For the sake of simplification, let us set α_i to 1. In this example, by setting $\alpha_i = 1$, we mean that there is always a relationship between θ_t and $P(Y_{it})$ controlling for X_t. We will consider four cases: where the direct and indirect effects are (1) of the same sign, (2) of differing signs, (3) the case when there is no indirect effect, and (4) the case when there is no direct effect. By examining the difference between these cases, it is clear that it is beneficial to consider both the indirect effect and direct effect. The marginal differences on item responses across groups are a poor substitute for understanding the underlying pattern of effects.

Suppose $\delta\alpha_i = 3$, meaning that there is a cohort difference in θ_t of 3. Additionally suppose there is a positive direct effect, $\gamma_i^\beta = 3$. Just examining the effect of cohort on the item (as is done in Table 1), not considering the latent variable, would reveal that there is a large effect of cohort on the item. Some researchers may see this and assume that there is differential item functioning based on this alone, or this difference relative to other items. However, half of the difference can be attributed to the indirect effect of cohort on the item response through the latent variable. This means that there is a cohort difference in θ_t and this contributes to an observed cohort difference in $P(Y_{it})$. The portion of the difference, $\delta\alpha_i$, which is attributable to the latent variable is still of interest to the researcher and could be examined even in the presence of DIF.

Consider now if $\delta\alpha_i = -3$ meaning that the 2006 cohort has a lower arithmetic ability than the 1933/36 cohort. Additionally suppose that $\gamma_i^\beta = 3$, which means that for item i, individuals in the 2006 cohort shows a higher probability of endorsing the item compared to individuals in the 1933/36 cohort with the same 'true' arithmetic ability. Note that in this case, we will not see any marginal cohort differences in $P(Y_{it})$. Some researchers may interpret this lack of marginal cohort differences as indicating that there is no DIF. However, because the direct and indirect effects cancel out, there are no marginal differences across groups on the item responses, but the magnitude of DIF is still the same as the previous example. It is the indirect effect which has changed, showing the indirect effect and direct effect must both be taken into account to understand the marginal patterns across groups on item responses.

If $\delta\alpha_i = 3$ and $\gamma_i^\beta = 0$ then the 2006 cohort has a higher 'true' arithmetic ability, and individuals from each cohort with the same latent ability have the same probability of getting the item correct. This illustrates a situation where there is an impact but no DIF (Ackerman, 1992; Camilli, 1993). Examining the marginal cohort differences on item responses, a researcher might see that the two cohorts respond differently to this item. The researcher may conclude that this is attributable to DIF, but that would be a mistake. Once the latent variable is taken into account it is clear that cohort differences come into play because there are cohort differences on the latent variable. When cohort, or the independent variable which causes DIF, is of primary interest, cases like this will likely be correctly identified as not attributable to DIF. However, if the effect of cohort on the latent variable were secondary, for example if the researcher were studying differences in arithmetic ability across different ages, cohort's effect on the latent variable could be overlooked, and the researcher may attribute cohort differences on item responses to DIF rather than to cohort differences on the latent variable.

Finally, consider the case where $\delta\alpha_i = 0$ and $\gamma_i^\beta = 3$. This means that there is no overall cohort difference in the 'true' arithmetic ability, but for item i individuals in the 2006 cohort show a higher probability of endorsing the item compared to individuals in the 1933/36 cohort with the same 'true' arithmetic ability. Note that in this case, we will observe a marginal cohort difference in $P(Y_{it})$, which is just the same as the third example. But the reason for this marginal group difference is attributable to DIF rather than the indirect effect. These are cases which would need to be correctly identified as DIF, and that is only done by examining both the direct and indirect effects. The marginal differences on item responses across groups are not informative enough to make a decision about DIF.

Consider the examples of DIF analysis for Item 5 and 10. Each item had marginal group differences on the item response probabilities which deviated from the rest of the items. However, we saw that Item 5 suffered from uniform DIF but Item 10 did not. These items seemed similar when examining marginal group differences on item response probabilities. However, the mechanism by which those marginal group differences arose are quite different.

The above exercise clearly explains why it is important to take into account potential group differences in the latent trait θ_t in investigating uniform DIF. Importantly, the indirect effect and direct effect are equally scaled and thus are directly comparable. This is what makes this exercise useful. Alternatively, researchers might be temped to compare α_i to γ_i^β. This comparison does not take into account the size of the effect of the independent variable on the latent variable, δ and thus these two effects would not be comparable as they are not on the same scale.

3.3 Symmetry of Moderation

The model expressed in Equation 6 allows the effect of θ_t on $P(Y_{it})$ to be moderated by X_t. Equation 6 can be re-written in an alternative way as follows:

$$\begin{aligned} P(Y_{it}|\theta_t, X_t) &= g^{-1}((\alpha_i + \gamma_i^\alpha X_t)\theta_t + \gamma_i^\beta X_t + \beta_i), &(8)\\ &= g^{-1}(\alpha_i\theta_t + \underbrace{(\gamma_i^\beta + \gamma_i^\alpha\theta_t)}_{\text{direct effect}=\gamma_i^{\beta *}} X_t + \beta_i). &(9)\end{aligned}$$

Note that Equation 8 shows that the effect of θ_t on $P(Y_{it})$ is moderated by X_t (see the first term), while Equation 9 specifies that the effect of X_t on $P(Y_{it})$ is moderated by θ_t (see the second term). Equations 8 and 9 are mathematically equivalent, thus statistical evidence that $\gamma_i^\alpha \neq 0$ suggests that X_t moderates the effect of θ_t on $P(Y_{it})$ *or* that θ_t moderates the effect of X_t on $P(Y_{it})$. There is no statistical way to distinguish between these two types of moderation. This is the symmetry property of moderation (Hayes, 2018; Hayes & Matthes, 2009).

Equation 9 can be useful for conceptualizing non-uniform DIF. The coefficient for X_t, is a function of θ_t. Recall that testing the coefficient of X_t in Equation 2 (i.e., the direct effect of X_t on $P(Y_{it})$), corresponded to testing a uniform DIF (discussed in Section 2.1). If the parameter γ_i^α takes a non-zero value (i.e., if non-uniform DIF exists for item i), this implies that the direct effect depends on the value of θ_t. In other words, group differences on $P(Y_{it})$ are not constant on the θ_t continuum.

An important implication is that in the presence of non-uniform DIF (or with a significant γ_i^α parameter value), the coefficient γ_i^β loses its interpretation as "uniform DIF" because it is no longer true that "group difference in the endorsement probability is constant over the latent continuum" (Woods & Grimm, 2011). Since $\gamma_i^{\beta^*} = \gamma_i^\beta$ holds only when $\theta_t = 0$ (if $\gamma_i^\alpha \neq 0$), the coefficient γ_i^β indicates the group difference on $P(Y_{it})$ when $\theta_t = 0$. In the arithmetic example, a significant $\gamma_i^\beta > 0$ means that individuals in the 2006 cohort with a 'true' arithmetic ability of zero (population mean) have a higher probability of endorsing item i compared to individuals from the 1933/36 cohort with the same arithmetic ability. It is important to interpret the γ_i^β coefficient correctly in the presence of non-uniform DIF.

3.4 Probing Conditional Effects

We have discussed that non-uniform DIF can be understood as an interaction between the latent variable of interest and an external grouping factor in predicting the probability of response on a given item, indicated by the coefficient γ_i^α. A non-zero γ_i^α value indicates that the group difference on $P(Y_{it})$ (or the direct effect) is not constant across the mediator θ_t. The interaction effect in DIF studies implies that the item characteristic curves for different groups cross at a point of the θ_t continuum. See Figure 1 for an illustration of different types of non-uniform DIF (Panels b, c, and d) in comparison to a zero non-uniform DIF case (Panel a).

In moderation analysis, once a significant moderation effect is found, researchers often apply probing methods (e.g., Hayes & Matthes, 2009; Spiller, Fitzimons, Lynch Jr., & McClelland, 2013). These methods can be used to understand where the group-specific item characteristic curves cross. Specifically, we can solve for the point along θ_t^* where cohort differences on $P(Y_ip)$ are estimated to be zero for item i by setting $\gamma_i^{\beta^*}$ to zero as follows:

$$\begin{aligned} \gamma_i^{\beta^*} &= \gamma_i^\beta + \gamma_i^\alpha \theta_t^* = 0, \\ \gamma_i^\alpha \theta_t^* &= -\gamma_i^\beta, \\ \theta_t^* &= -\frac{\gamma_i^\beta}{\gamma_i^\alpha}. \end{aligned}$$

It is useful to know the point of θ_t^* where the group-specific item characteristic curves cross because this point informs us that one group has a higher probability of endorsing the item of interest, compared to the other groups whose latent trait value is $\theta_t < \theta_t^*$; however, the opposite is true for those people with $\theta_t \geq \theta_t^*$. See Figures 1b and 1c for the non-uniform DIF cases with different θ_t^* locations. If the value of θ_t^* is very small (e.g., smaller than -5) or very large (e.g., larger than 5), the group-specific item characteristic curves may look parallel (no crossing) within the θ_t range from -5 to 5. See Figure 1d for such an example. It is also possible to make an inference about the point θ_t^* after computing its standard error using the a delta method.

4 Discussion

In this paper, we have discussed how the MIMIC models for studying uniform and non-uniform DIF can be conceptualized within the mediation and moderation framework. Specifically, we showed how estimating and testing a direct effect in a mediation model aligns with a test of uniform DIF and how a test of an interaction effect in a moderated mediation model aligns with a test of non-uniform DIF. A benefit of conceptualizing DIF within the mediation and moderation framework is that we can apply useful ideas and methods from mediation and moderation literature to improve our understanding of DIF. Typical DIF studies are often exploratory and researchers who wish to study DIF tend to act as if they have no prior information about what type of DIF they might expect. However, understanding DIF in the mediation and moderation context may help researchers apply their substantive knowledge in such a way as to develop more directed hypotheses about DIF for particular items. By developing specific hypotheses about DIF, researchers can transition into a more confirmatory study of DIF.

An additional benefit of conceptualizing DIF within the mediation and moderation framework is that we clarified that the coefficient γ_i^β loses its original interpretation in the presence of non-uniform DIF (or with non-zero γ_i^α). Further, we discussed potential usefulness of examining indirect effects in DIF research.

Throughout this note we have discussed the MIMIC models for exploring differential item functioning, however the ideas in this manuscript generalize to other models as well. The MIMIC model is equivalent to the two parameter logistic model (B. O. Muthèn et al., 1991; MacIntosh & Hashim, 2003). When a one parameter logistic model is desired, the MIMIC model can be adjusted so the relationship between θ_t and $P(Y_{it})$ is constrained to 1 for all items. Finch (2005) found that MIMIC models perform comparably to other uniform DIF detection methods for data generated from a 3PL model, except when the scale is short (20 items or less). Shih and Wang (2009) showed that data generated using a 3PL model can be effectively analyzed for uniform and non-uniform DIF using MIMIC models. However there is lower power to detect DIF for data generated from a 3PL compared to a 2PL. Constructing a MIMIC model which incorporates the guessing parameter is not straightforward. However, we believe it is still useful to discuss uniform and non-uniform DIF with a 3PL model using the concepts of mediation and moderated mediation.

Throughout this paper we described the analyses using a general link function, but for the data example we used a logistic link function. When a probit link is used we can construct a normal-ogive MIMIC model, which is equivalent to a normal ogive logistic model. B. O. Muthèn (1985) showed how to derive the parameters for the normal-ogive IRT model from a normal-ogive MIMIC model. MacIntosh and Hashim (2003) followed up on this work deriving the standard errors for the parameters for the normal-ogive IRT model from the standard errors in the normal ogive MIMIC model. The MIMIC model can be used for either logistic or normal-ogive link functions.

In Section 1.2 we discussed previous research on selecting anchor items and scale purification using the MIMIC model. These methods, described by Wang et al. (2009) and Shih and Wang (2009) have only investigated detection methods for uniform DIF. The non-uniform DIF MIMIC model is slightly more recent (Woods & Grimm, 2011), and no research has yet to explore how best to go about selecting anchor items or conducting scale purification using MIMIC models for non-uniform DIF. Future research should examine how best to do this type of analysis, particularly when some items have uniform DIF and others have non-uniform DIF.

We are not the first to discuss some of the connections between mediation analysis and DIF. Cheng et al. (2016) proposed applying mediation analysis to understand how uniform DIF occurs. Specifically, Cheng et al. (2016) proposed to introduce additional mediators in a MIMIC model to explain the process of uniform DIF; however, the authors did not acknowledge that a MIMIC model could already be seen as a mediation model as we discussed in this note. Adding extra mediators in a MIMIC model for uniform DIF as suggested by Cheng et al. (2016) would result in extending a single mediator model to a multiple mediator model in our framework. The additional mediators can be used to test whether the uniform DIF (or a direct effect of X_t on $P(Y_{it})$) may be accounted for by the additional mediators. We have discussed the possibility of including additional mediators in Section 3.1 in the context of interpreting null direct effects.

An additional benefit of MIMIC models is that multiple independent variables can be included in the model, as was briefly discussed in Section 3.1. Including multiple indepedent variables would allow the researcher to estimate indirect and direct effects for each independent variable. However, it is important to remember that indirect and direct effects are scaled by the independent variable (i.e., they are interpreting with respect to a one unit change in the covariate). This means that direct effects through different independent variables may not be directly comparable in terms of magnitude.

Finally, we would like to mention that though conceptualizing uniform and non-uniform DIF MIMIC models within a mediation and moderation framework can be very useful, we caution against using this framework to make causal inferences without thoroughly investigating the assumptions needed to do so. There is a growing literature on how to make valid causal inferences, which may be additionally complicated by the use of a latent variable. Researchers interested in making causal inferences should consult the literature on causal mediation analysis (e.g., Imai, Kelle, & Tingley, 2010; Robins & Greenland, 1992; Coffman, MacKinnon, Zhu, & Ghosh, 2016).

5 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*(1), 67 – 91.

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*, 51.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397 – 413). Hillsdale, NJ: Lawrence Earlbaum.

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, *76*(1), 43 – 63.

Coffman, D. L., MacKinnon, D. P., Zhu, Y., & Ghosh, D. (2016). A comparison of potential outcome approaches for assessing causal mediation. In H. He, P. Wu, & D.-G. Chen (Eds.), *Statistical Causal Inferences and Their Applications in Public Health Research* (pp. 263 – 293). New York: Springer.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278 – 295.

Haggerty, M. E., Terman, L. M., Thorndike, E. L., Whipple, G. M., & Yerkes, R. M. (1921). *National intelligence tests: Manual of directions.* New York: World Book Company.

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1 – 22.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). New York, NY: Guilford Press.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*(3), 924 – 936.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Brown (Eds.), *Test validity* (pp. 129 – 145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Imai, K., Kelle, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309 – 334.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409 – 426.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331 – 358.

Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicator multiple cases models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545 – 569.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Earlbaum.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372 – 379.

Magis, D., Bèland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847 – 863.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statitics*, *7*(2), 105 – 118.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297 – 334.

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*(6), 780 – 790.

Must, O., & Must, A. (2014). Data from "Changes in test-taking patterns over time" concerning the Flynn Effect in Estonia. *Jounal of Open Psychology Data*, *2*(1), e2.

Muthèn, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*(2), 121 – 132.

Muthèn, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557 – 585.

Muthèn, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1 – 22.

Muthèn, L. K., & Muthèn, B. O. (1998 – 2011). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthèn & Muthèn.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257 – 274.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197 – 207.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143 - 155.

Rucker, D. D., Preacher, K. J., & Tormala, Z. L. (2011). Mediation analysis in social psychology: Current practice and new recommendations. *Personality and Social Psychology Compass*, *5/6*, 359 – 371.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTT as well as item bias/DIF. *Psychometrika*, *58*, 159 – 194.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184 – 199.

Spiller, S. A., Fitzimons, G. J., Lynch Jr., J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, *50*, 277 – 288.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB – a procedure to investigate DIF when a test is intentionally multidimensional. *Applied Psychological Measurement*, *21*, 195 – 213.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361 – 370.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713 – 731.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1 – 27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339 – 361.

Yun, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California Los Angeles.

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*(2), 223 – 233.

Table 1

Estonian National Intelligence Test: Arithmetic

Item	Item Text English Translation [1]	Proportion Correct	
		1933/36	2006
1	How much is half of 8 minutes?	0.9273	0.9651
2	How many dimes in 6 nickels?	0.7700	0.8651
3	Mari buys an apple for 4 cents and a cake for 5 cents. How many cents does he have to pay?	0.9400	0.9759
4	Vilu had 8 chickpeas and sold 3. How many are left?	0.9831	0.9857
5	How much longer is 12 yards than a meter?	0.7526	0.6625
6	How many chairs are in 9 rooms if there are 40 chairs in each room?	0.8467	0.8698
7	A gift costs 96 cents between 4 girls. How much should each girl pay?	0.8795	0.8929
8	How much do 12 sweets cost if 3 sweets costs one cent?	0.5393	0.5886
9	How many square centimeters is a card with a length of 5 and a width of 3 cm?	0.9510	0.8482
10	A man bought a plot of land for 100 kroons and sold it for 120 kroons. He profited 5 kroons per acre. How many acres was the plot of land?	0.5258	0.5051
11	How many times does $1\frac{1}{2}$ need to be added to 6 to get 15?	0.5161	0.4981
12	Half a kilo of seeds costs 8 kroons. How many seeds can you buy with 50 kroons?	0.1787	0.0721
13	It is necessary to carry 56 kilograms of equipment to the camp. A, B, and C distribute the equipment among themselves so that 3 parts are to A, 2 part B and 2 parts C. How many kilos did A need to wear?	0.4491	0.4194
14	How many bulls does a hunter have to shoot to hit the mark 100 times when it hits on 40% of the shots?	0.1859	0.2425
15	How many times heavier is $\frac{1}{2}$ of a load weighing one and a half tons than a half-ton load?	0.1140	0.1266
16	The pocket watch was set correctly at 12 noon on Wednesday. At 6 o'clock the next night, it was 15 seconds ahead. How much does the clock go for half an hour?	0.0426	0.0692

[1] Original items are from Haggerty (1921) Scale A, Form 2, Edition 2. Items were translated to English for ease of understanding in this manuscript. Items were administered in Estonian.

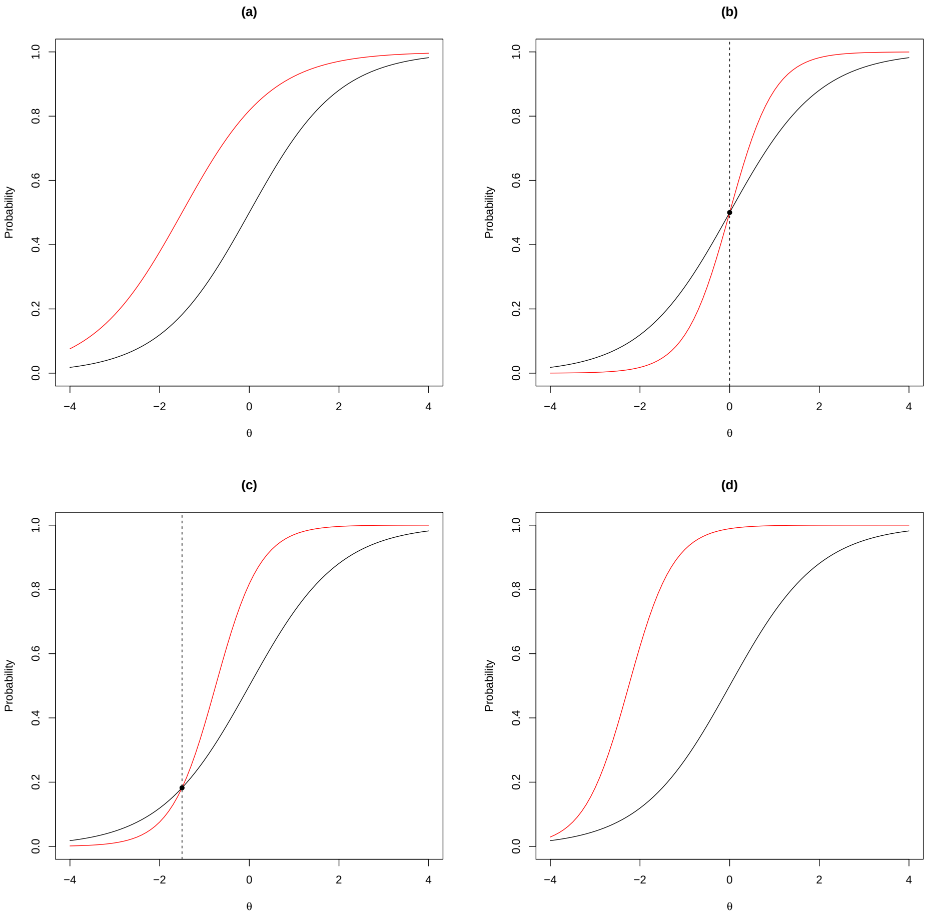


Figure 1. Item characteristic curves when (a) $\gamma_i^{\beta} = 1.5$ and $\gamma_i^{\alpha} = 0$, (b) $\gamma_i^{\beta} = 0$ and $\gamma_i^{\alpha} = 1$, (c) $\gamma_i^{\beta} = 1.5$ and $\gamma_i^{\alpha} = 1$, and (d) $\gamma_i^{\beta} = 4.5$ and $\gamma_i^{\alpha} = 1$. Panel (a) illustrates uniform DIF, while (b), (c), and (d) illustrate three different types of non-uniform DIF.

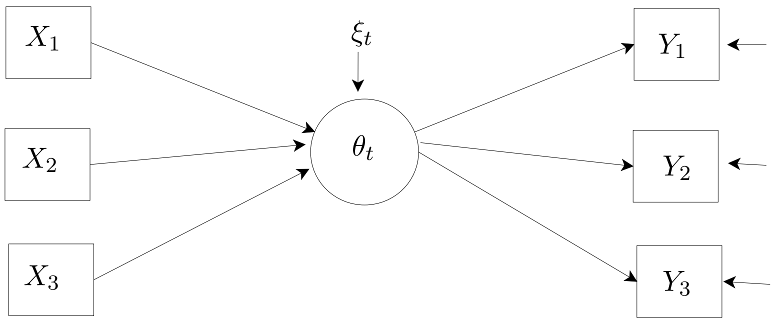
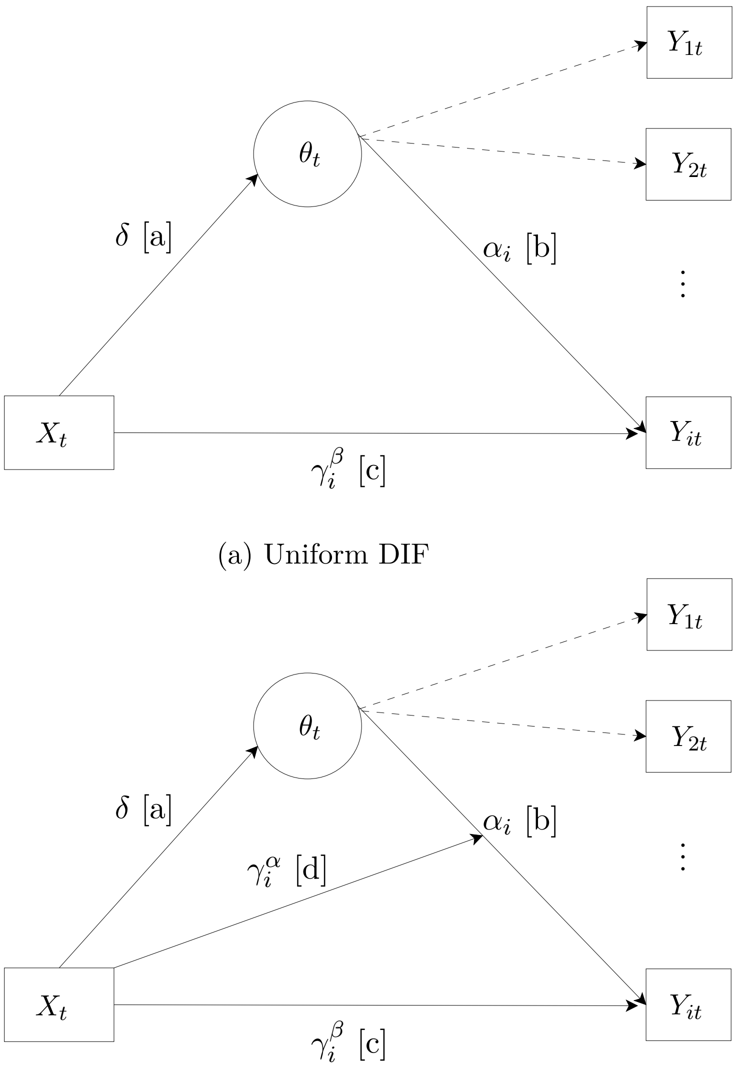


Figure 2. An example of MIMIC model with three observed indicators (Y_1, Y_2, Y_3) and three observed causes (X_1, X_2, X_3) for the latent variable θ_t (with ξ_t being a residual).



(a) Uniform DIF

(b) Uniform and Non-uniform DIF

Figure 3. (a) A MIMIC model for uniform DIF as a mediation model and (b) a MIMIC model for non-uniform DIF as a moderated mediation model.

Appendix A

Mplus Code for Must and Must (2013) Analysis: Item 5, Uniform DIF

```
TITLE: MIMIC DIF model Item 5 Uniform DIF;
DATA: FILE IS FEarithdat.csv;

!Original grouping coding was 1/2, recodes to 0/1;
DEFINE: group = group - 1;

!Group codes cohort 1933/36 vs 2006,  Q1-Q16 are arithmetic items;
VARIABLE:NAMES ARE group Q1-Q16;

!Sets all items to be categorical;
CATEGORICAL ARE Q1-Q16;

!Missing values coded as -9;
MISSING = ALL(-9);

!Maximum likelihood estimation with robust standard errors;
ANALYSIS: ESTIMATOR = MLR;

!Latent factor is indicated by 16 measured variables;
MODEL: f by Q1 - Q16* ;

!Factor variances set to 1;
f@1;

!Group predicts latent factor;
f on group;

!Group predicts Question 5;
Q5 on group;
```

Appendix B

Mplus Code for Must and Must (2013) Analysis: Item 10, Non-Uniform DIF

```
TITLE: MIMIC DIF model Item 10 Non-Uniform DIF;
DATA: FILE IS FEarithdat.csv;

!Original grouping coding was 1/2, recodes to 0/1;
DEFINE: group = group - 1;

VARIABLE:
!Group codes cohort 1933/36 vs 2006, Q1-Q16 are arithmetic items;
NAMES ARE group Q1-Q16;

!Sets all items to be categorical;
CATEGORICAL ARE Q1-Q16;

!Missing values coded as -9;
MISSING = ALL(-9);

!Maximum likelihood estimation with robust standard errors;
ANALYSIS: ESTIMATOR = MLR;

!Random type required for latent variable interactions
!allows different variances for f in each cohort;
type=random;

!Latent factor is indicated by 16 measured variables;
!Name the loading for Q10 "alpha";
MODEL: f by Q1 - Q9*
Q10 (alpha)
Q11 - Q16;

!Factor variances set to 1;
f@1;

!Group predicts latent factor;
f on group;

!Define interaction between latent factor and group;
```

```
interact2 | f xwith group ;

!Group and interaction predict Question 10;
!Name weight for interaction "gammaalpha";
Q10 on group
interact2 (gammaalpha);

!create new variable for discrimination parameter in 2006 cohort
MODEL CONSTRAINT: new (disc2006);
disc2006 = alpha + gammaalpha;
```