AN INVESTIGATION OF DIFFERENTIAL ITEM FUNCTIONING IN DIAGNOSTIC CLASSIFICATION MODELS

by

SELAY ZOR

(Under the Direction of Matthew James Madison)

ABSTRACT

Diagnostic classification models (DCMs) are statistical models that provide diagnostic information about the mastery state of examinees' knowledge component or attributes. DCMs classify examinees based on these specified attributes, providing fine-grained and multidimensional diagnostic information. This information can be used by educators in designing targeted instructional interventions to enhance student performance. In DCMs, the inclusion of DIF analysis and effect size measures holds significant importance. DIF analysis enables the identification of potential biases and ensures test fairness in diagnostic assessments by detecting items that function differently across different groups. This process enhances the accuracy of diagnostic classifications by identifying items that may unfairly advantage or disadvantage certain groups. Furthermore, the incorporation of effect size measures in DIF analysis provides valuable insight into the practical significance of DIF. The effect size measures facilitate the interpretation of DIF results and aid in decision-making regarding the treatment of DIF items. This dissertation consists of two studies that address DIF analysis in a general DCM framework. The first study focuses on evaluating the performance of the Wald DIF detection method using the LCDM as a general framework under various conditions. Through simulation conditions, we

investigate the performance of Wald DIF under the LCDM model and compare its performance with the likelihood ratio test (LRT) for DIF detection. In the second study, we investigate the practical significance of DIF in DCMs. We examine criteria based on the degree to which DIF items impact DCM classifications and reliability. A simulation study is conducted to investigate the effects of DIF on classifications, and the results are compared with the unsigned area (UA) effect size measure to provide guidelines for flagging DIF items.

INDEX WORDS: diagnostic classification models, cognitive diagnosis models, differential item functioning, Wald test, DIF effect size measures

AN INVESTIGATION OF DIFFERENTIAL ITEM FUNCTIONING IN DIAGNOSTIC CLASSIFICATION MODELS

by

SELAY ZOR

B.S., Balikesir University, Turkey, 2013

M.A., The University of Georgia, 2018

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2023

AN INVESTIGATION OF DIFFERENTIAL ITEM FUNCTIONING IN DIAGNOSTIC CLASSIFICATION MODELS

by

SELAY ZOR

Major Professor:	Matthew James Madison
Committee:	Allan S. Cohen
	Shiyu Wang
	Jaxk Reeves

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2023

DEDICATION

To my parents, Leyla and Hasan Zor, my sister, Sumeyra Kostereli, and my brother, Furkan Zor, for their endless love and support.

ACKNOWLEDGEMENTS

I first would like to thank my advisor and committee chair, Dr. Matthew Madison, for his exceptional expertise, guidance, immense patience, and constant encouragement throughout the entire process of writing this dissertation. He has not only been an incredible advisor but also an inspiring scholar, setting a remarkable example as a collaborative and supportive researcher. I am profoundly grateful for his invaluable guidance and unwavering support in every meeting, which also alleviated my stress. Without his exceptional guidance and support, this dissertation would not have been possible.

I would like to thank my committee members, Drs. Allan Cohen, Shiyu Wang, and Jaxk Reeves, for their invaluable suggestions and constructive feedback. I am grateful for the time they dedicated to reviewing my work and providing insightful comments.

I would also like to thank my first graduate advisor, Laine Bradshaw, for her invaluable guidance and encouragement during the early years of my journey in graduate school. Her support played a significant role in shaping my academic growth and research trajectory.

I want to thank my friends and QM friends for their companionship throughout this journey. Their encouragement and uplifting presence have provided me with tremendous motivation. My besties, thank you for your endless support and keeping me nourished during the challenging moments of dissertation writing.

Lastly, but most importantly, I want to thank my beloved parents, Leyla and Hasan Zor, my sister, Sumeyra Kostereli, and my brother, Furkan Zor. Their unwavering support from the very beginning has been invaluable. They have always been my pillars of strength, continuously

encouraging me and helping me overcome every challenge with their boundless love. A special note of gratitude goes to my sister and brother-in-law, Sumeyra and Ziya Kostereli, for visiting me during this time and creating unforgettable memories. Lastly, I want to express my heartfelt appreciation to my beautiful niece, Hale, for patiently waiting for her auntie to return and create more cherished moments together. I deeply regret not being able to witness the precious years of your infancy, but I am excited for the countless beautiful memories that lie ahead.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTODUCTION AND LITERATURE REVIEW

Diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010), also known as cognitive diagnostic models (CDMs; e.g., Leighton & Gierl, 2007), have emerged as powerful tools in educational research and assessment. DCMs are confirmatory multidimensional latent-variable models that provide diagnostic information about the mastery state of examinees' knowledge or skills (Rupp & Templin, 2008). By evaluating examinees' mastery status on each targeted skill or knowledge component, referred as *attribute*, DCMs classify examinees into attribute profiles, providing fine-grained and multidimensional diagnostic information. This information can be used by educators in designing targeted instructional interventions to enhance student performance. Despite the growing attention and utilization of DCMs, further research is needed to address critical psychometric considerations that underpin validity and fairness. One important aspect that requires investigation is differential item functioning (DIF) within the DCM framework. Understanding DIF and its potential influence on classifications is crucial to ensure the accuracy of the diagnostic information provided. This dissertation focusses on investigating DIF detection and practical significance of DIF in a general DCM framework. This chapter presents an introduction of general DCMs, followed by an overview of DIF and effect size measures in traditional framework and DCM framework.

Diagnostic Classification Models

Diagnostic classification models are confirmatory multidimensional latent variable models that can provide diagnostic information regarding the mastery state of knowledge components (Rupp et al., 2010). Unlike traditional models that assume continuous latent variables, DCMs assume discrete latent variables. These variables are referred to as *attributes* in the literature. A diagnostic classification model is a statistical model used to analyze educational or psychological data that aims to classify examinees into distinct latent classes or mastery states based on their performance on a set of test items. These latent classes represent different levels of proficiency or mastery on specific knowledge components or attributes. DCMs provide a comprehensive framework for understanding and evaluating an individual's proficiency in multiple attributes, enabling educators and researchers to gain valuable insights into an examinee's strengths, weaknesses, and specific areas of instructional need (Rupp & Templin, 2008). Moreover, a key advantage of DCMs lies in their ability to provide reliable estimates of examinees' latent attributes with fewer items compared to IRT models (Templin & Bradshaw, 2013).

As the confirmatory nature of DCMs points out, the latent classes representing different knowledge components are defined, and the attribute loading structure is explicitly specified in a Q-matrix. Once the attributes and latent classes are set, the probability of an examinee answering an item correctly is modeled as a function of their *attribute profile*, which indicates their mastery levels in latent classes. For a test measures *A* attributes, DCMs classify examinees into one of 2^A attribute profiles based on their item responses. The attribute profile for examinee *e* is an *A* length vector, $\boldsymbol{\alpha_e} = [\alpha_1, \alpha_2, \ldots, \alpha_A]$, where for the binary level of attribute mastery 1 indicates mastery and 0 indicates nonmastery of the attribute. For example, let's consider an attribute

profile [1,0,1]. This profile suggested that the examinee has mastered Attributes 1 and 3, as indicated by the value of 1, while not mastered Attribute 2, as indicated by the value of 0. An attribute profile is assumed to provide information about examinees' strengths and weaknesses in certain attributes. The Q-matrix (Tatsuoka, 1990) defines the relationship between items and attributes by indicating which attributes are measured by each item. The Q-matrix is determined a priori and provides information about the item-attribute associations, denoted as $\boldsymbol{q_i} = [q_{i1}, q_{i2}, \ldots, q_{iA}]$. Each element in the Q-matrix, denoted as q_{ia}, indicates whether attribute *a* is measure by item *i*. Specifically, q_{ia} takes a value of 1 if the item measures attribute *a*, and 0 if the item does not measure attribute *a*. An item can be designed to measure one attribute exclusively or multiple attributes.

Within the framework of latent class models (LCMs; Lazarsfeld & Henry, 1968), DCMs adopt a general latent class parameterization. DCMs consist of two key components: the structural model and the measurement model. The structural component indicates the distribution of examinees across latent classes, indicating the proportion of examinees within each class. The measurement component describes the response probabilities for each item within each latent class. For an examinee *e*, the item response probability is:

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}^{x_{ei}} (1 - \pi_{ic})^{1-x_{ei}} \quad (1)$$

where $P(\boldsymbol{X_e} = \boldsymbol{x_e})$ is the probability for the observed response pattern $\boldsymbol{x_e}$, and the structural component represented by the summation part, while the measurement component represented by the product part. In Equation 1, v_c is the proportion of examinees who have mastered the attributes required by the class *c*; x_{ei} is the observed response to Item *i* by an examinee *e*; and π_{ic} is the probability of answering Item *i* correctly by an examinee in latent class *c*. The product

implies that item responses are independent within a latent class as a consequence of the local independence. The structural component represents latent class membership probabilities which provides the base-rate proportion of examinees in Class *c*.

A large number of DCMs have been developed based on the ways they parameterize the response probabilities. Those models can be categorized as compensatory and noncompensatory models. In compensatory models, mastering a subset of the attributes required by the item can compensate for the non-mastery of the remaining attributes. In noncompensatory models, however, all required attributes need to be mastered to produce a correct response. The deterministic input, noisy and gate model (DINA; Junker & Sijtsma, 2001) is an example of a noncompensatory model, and the deterministic input noisy or gate model (DINO; Templin & Henson, 2006) is an example of a compensatory model. While mastering an additional attribute does not increase the probability of giving a correct answer in a noncompensatory model, compensatory models allow the increase in the probability as mastering additional attributes. The log-linear diagnosis model (LCDM; Henson, Templin, & Willse, 2009), described in the next section, provides a general framework for DCMs. By placing constraints on the LCDM parameters, both compensatory and noncompensatory core DCMs can be specified.

The Log-linear Cognitive Diagnosis Model

The LCDM is a general DCM that provides a framework to model the relationship between item responses and attributes. A key feature of the model is that core DCMs can be represented when some parameters are included or constrained in the LCDM, which makes these models nested within the LCDM. To explain the item response function, assume an item measures Attribute 1 and Attribute 4, Q-matrix entries are q_{i1}= 1 and q_{i4}= 1. The LCDM item response function is

$$P(\boldsymbol{X_{ie}} = \boldsymbol{x_e} \mid \boldsymbol{\alpha_e}) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(4)}\alpha_{e4} + \lambda_{i,2,(1,4)}\alpha_{e1}\alpha_{e4}\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(4)}\alpha_{e4} + \lambda_{i,2,(1,4)}\alpha_{e1}\alpha_{e4}\right)} \quad (2)$$

In Equation 2, X_{ie} is the observed response to item *i* by examinee *e* whose attribute profile is α_e. The LCDM item parameters are akin to a dummy-coded ANOVA model with an intercept, a main effect for each attribute, and interaction term(s) for combinations of attributes. The subscript *i* on the λ parameters represents the item *i*. The second subscript on the λ parameters represents the parameter level: The subscript 0 is for intercept parameters, 1 is for main effects, 2 is for two-way interactions, 3 is for three-way interactions, etc. The third subscript is in parentheses and represents the attributes to which the main effects or interactions apply. In Equation 1, $\lambda_{i,0}$ is the intercept representing the log-odds of a correct response for examinees who have not mastered either Attribute 1 or Attribute 4. $\lambda_{i,1,(1)}$ and $\lambda_{i,1,(4)}$ are the main effects that represent the increase in log-odds for examinees who possess either Attribute 1 or Attribute 4. $\lambda_{i,2,(1,4)}$ is the two-way interaction term that indicates the change in log-odds of a correct response when examinees have mastered both of the attributes.

While the item measures two attributes in the example above, the LCDM can include more that two attributes resulting in additional main effects and interactions. In the LCDM framework, the probability of a correct response to item *i* is conditional on an examinee *e*'s attribute profile $\alpha_e = \boldsymbol{\alpha_c}$. The general form of LCDM item response function is

$$P(X_{ie} = 1 \mid \boldsymbol{\alpha}_c) = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i})\right)}{1 + \exp\left(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i})\right)} \quad (3)$$

In Equation 3, $\boldsymbol{q_i}$ represents the Q-matrix entries for item *i*, and $\lambda_{i,0}$ is the intercept parameter as described above. $\boldsymbol{\lambda_i}$ is a vector of size $(2^A - 1) \times 1$ containing main effect and interaction parameters for item *i*, *A* is the number of attributes; and $\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q}_i)$ is a vector of size $(2^A - 1) \times 1$ representing a set of linear combinations of $\boldsymbol{\alpha_c}$ and $\boldsymbol{q_i}$. $\boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_i)$ is written as:

$$\boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q}_i) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ca}\, q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a}^{A} \lambda_{i,2,(a,b)} \alpha_{ca} \alpha_{cb} q_{ia} q_{ib} + \cdots \quad (4)$$

where $\lambda_{i,1,(a)}$ represents the main effect of Attribute *a* on the item *i*, and $\lambda_{i,2,(a,b)}$ represents the two-way interaction effect of Attribute *a* and Attribute *b*. The right side of the equation includes all main effect parameters and all possible interaction parameters. Item parameters are present if the linear combination of $\boldsymbol{\alpha_c}$ and $\boldsymbol{q_i}$ equals 1. For main effects, this only occurs when an examinee has mastered the attribute ($\alpha_{ea} = 1$) and item measures the attribute ($q_{ia} = 1$). For two-way interaction terms, similarly, both attributes are needed to be mastered by an examinee ($\alpha_{ea} = 1, \alpha_{eb} = 1$) and item measures the attributes ($q_{ia} = 1, q_{ib} = 1$). The LCDM models the item responses in a similar way with ANOVA; the attributes and attribute mastery in the LCDM serve as factors and levels of the factors in an ANAOVA, respectively. Constraints are placed on the main effect parameters and interaction terms so that examinees' correct response probabilities increase for DCMs represented by the LCDM (Rupp et al., 2010).

Differential Item Functioning

Test fairness is a fundamental aspect of conducting group comparisons and ensuring the validity of assessments, particularly when examining differences based on gender, ethnicities, culture, or treatment conditions. To achieve test fairness, it is essential to detect and prevent any form of unfairness throughout the entire testing process, including test design, development, administration, and scoring (Camilli, 2006; Dorans & Cook, 2016). Differential item functioning (DIF) analysis plays a crucial role in addressing test fairness by identifying potentially biased items in a test. DIF procedures assess whether examinees from different subgroups, who possess the same underlying ability or trait, have different probabilities of endorsing an item (Angoff, 1993; Camilli, 2006). By identifying items that function differently across groups, DIF analysis

helps to minimize the impact of factors unrelated to the construct being measured (Sireci & Rios, 2013; Zumbo, 2007). In the absence of bias, responses to an item reflect the level of the underlying trait being measured. However, the presence of item bias indicates that responses to the item are influenced by factors other than the underlying trait level. Thus, biased items systemically advantage or disadvantage a specific subgroup because of the factors irrelevant to the intended construct. By addressing DIF, the fairness and validity of the test can be enhanced, ensuring that an item is unbiased and measure the same construct across groups.

Differential item functioning is defined as an item functioning differently in different groups after matching examinees on a proficiency measure (Angoff, 1993). In classical test theory (CTT), an individual's latent trait is predicted based on the observed score. More specifically, the observed score serves as the theta estimate in CTT framework. Therefore, the observed score has been used as a matching criterion for investigating DIF in the context of CTT. On the other hand, based on the latent variable approach, the ability estimates have been used as a matching criterion in item response theory (IRT). In order to explain the relationship between the latent ability measured by the test and the probability of a response to an item, IRT models establish a link between item properties, examinees, and the underlying ability. The item characteristic curve (ICC) demonstrates this information for a particular IRT model. Having said that, IRT provides a framework to study DIF by comparing ICCs. DIF occurs when the ICCs differ for the studied groups, and one way to detect DIF is to compare the ICCs from different groups or equally the item parameters. More generally, in an IRT framework, DIF occurs if the probability of answering an item correctly differs in one group compared to another group(s) for examinees who have the same ability (Cohen, Kim, & Baker, 1993; Hambleton & Rogers, 1989; Holland & Thayer, 1988; Swaminathan & Rogers, 1990). While IRT models provide continuous latent

ability estimate for each examinee and locate them on a latent continuum, it is one of the key characteristics of DCMs to provide latent attribute profiles to the examinees that indicates whether the examinee has mastered or not mastered the measured attributes in an assessment. Due to the multidimensionality assumption of DCMs, we need to redefine DIF in the DCM framework.

Most DIF detection methods use a unidimensional raw score, or latent ability estimate as matching criteria. Under a multidimensional test condition, however, using unidimensional criterion for matching the groups may result in flagging more items as DIF (Ackerman, 1992). In order to help to reduce the inflated estimates of DIF, more appropriate proficiency measures as matching variable are required in DCMs. Instead of a single ability estimate, DCMs provide latent attribute profiles that have been used as a matching criterion for DIF analysis under the DCM framework (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Zhang, 2006). Zhang (2006) used attribute mastery profiles as the matching criteria to detect DIF under the DINA model and found that the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) using attribute profiles as matching variable resulted in better Type I error and power rates than using total scores as matching variable. In this case, DIF in the context of DCMs is defined as the difference in the probability of a correct response in one group compared to another group (or groups) conditioned on the latent attribute profiles (Hou et al., 2014; Li & Wang, 2015).

Defining DIF in the DCM Framework

In DCM framework, an item is considered to demonstrate DIF when the probabilities of success on the item differ among examinees who possess the same latent attribute profile but belong to different groups (Hou et al., 2014; Li & Wang, 2015). In other words, the presence of

DIF indicates that the item behaves differently for examinees with the same attribute profile but different group affiliations.

Let $\Delta_{i\alpha_e}$ be DIF in item *i* for examinees with the attribute profile α_e. In DCM, DIF can be expressed as follows:

$$\Delta_{i\alpha_e} = P(X_i = 1|\, \alpha_e)_{\mathrm{F}} - P(X_i = 1|\, \alpha_e)_R \quad (5)$$

where $P(X_i = 1|\, \alpha_e)_{\mathrm{F}}$ and $P(X_i = 1|\, \alpha_e)_R$ is the probability of answering item *i* correctly for examinees who have the attribute profile α_e in the focal group and reference group, respectively. When $\Delta_{i\alpha_e} = 0$ for all latent attribute profiles, item *i* shows no DIF. On the other hand, when $\Delta_{i\alpha_e} > 0$, item *i* shows DIF that is against the reference group and when $\Delta_{i\alpha_e} < 0$, item *i* shows DIF that is against the focal group.

Similar to DIF in IRT, both uniform and nonuniform DIF can also be defined in DCMs (Hou et al., 2014). Uniform DIF refers to a situation where the probabilities of correctly answering an item are consistently lower or higher for one group, irrespective of the latent attribute profile. In other words, a uniform DIF is present when the item shows a systematic bias towards one group across all levels of attribute mastery, indicated by a consistent positive or negative value ($\Delta_{i\alpha_e}$) for all latent attribute profiles. On the other hand, nonuniform DIF occurs when the success probabilities of an item are lower for one group on certain attribute profiles and higher for other attribute mastery profiles. In this case, the sign of $\Delta_{i\alpha_e}$ varies depending on the attribute profiles.

In the LCDM model, each item parameter corresponds to the probability of success on item *i* for examinees with the attribute profile α_{e}. For the LCDM model, DIF can be formulated as the difference in the item parameters between the focal and reference groups. Item *i* exhibits DIF when

$$\Delta_{\lambda_{i,0}} = \left(\lambda_{i,0}\right)_F - \left(\lambda_{i,0}\right)_R \neq 0, \quad (6)$$

and

$$\Delta_{\boldsymbol{\lambda_i}} = (\boldsymbol{\lambda_i})_F - (\boldsymbol{\lambda_i})_R \neq 0. \quad (7)$$

where $\Delta_{\lambda_{i,0}}$ represents the difference in intercept parameters, and $\Delta_{\boldsymbol{\lambda_i}}$ represents the difference in main effect and possible interaction parameters for item *i*. In cases where DIF is in intercept parameter, uniform DIF occurs (i.e., $\Delta_{\lambda_{i,0}} < 0$ or $\Delta_{\lambda_{i,0}} > 0$). For example, when intercept parameter is larger for the reference group, the probabilities of correctly answering the item for examinees in reference group is higher compared to those in the focal group regardless of their attribute mastery profile, indicating the item favors the reference group (i.e., $\Delta_{\lambda_{i,0}} < 0$). Nonuniform DIF occurs when DIF is in the main effect and interaction parameters since the change in these parameters affect the item response probabilities for masters only (i.e., $\Delta_{\boldsymbol{\lambda_i}} < 0$ or $\Delta_{\boldsymbol{\lambda_i}} > 0$). For example, when main effects and possible interaction parameters for item *i* is larger for the reference group ($\Delta_{\boldsymbol{\lambda_i}} < 0$), the correct response probabilities of the masters in reference group are higher compared to those masters in the focal group, and no change for nonmasters. This indicates that the item favors the reference group for masters yet no change for the nonmasters. Another way to produce uniform and nonuniform DIF is when DIF is introduced to all parameters. Uniform DIF occurs in item *i* when both $\Delta_{\lambda_{i,0}}$ and $\Delta_{\boldsymbol{\lambda_i}}$ have the same signs, and having different signs would result in nonuniform DIF depending on the canceling effect.

DIF Detection and Effect Size Measures in Traditional Framework

It is especially crucial that DIF is addressed in order to identify any potential biased items in a test that are functioning differently among the different groups after the ability levels for the groups are equal. DIF analysis typically involves comparing two groups: the focal group, which

is the group of primary concern, and the reference group, which serves as a comparison group. Two primary approaches for DIF analysis exist: one based on observed scores and the other utilizing IRT models. In both approaches, the performance on a specific item is compared after matching examinees on the ability of interest. This enables the identification of items that may exhibit DIF, indicating potential biases in the test across different groups.

The selection of an appropriate method for investigating DIF is crucial for obtaining accurate and meaningful results. Researchers and practitioners have a range of options to choose from, and the decision is typically guided by various factors related to the specific research context. These factors include the number of groups being compared, characteristics of the data, the nature of the DIF (e.g., uniform and nonuniform), and the nature of the method itself (Svetina, Dai, and Wang, 2017). The choice of method should align with the specific research context and requirements to ensure accurate and reliable detection of DIF.

In studies investigating DIF, it is not only important to detect significant DIF but also to consider the effect size of the detected DIF. Effect size measures provide quantification of the magnitude and practical significance of DIF, enabling a more comprehensive understanding of the impact of DIF. Researchers have highlighted the significance of examining effect size measures alongside hypothesis testing to determine the practical importance of DIF (Jodoin & Gierl, 2001; Kirk, 1996). This is particularly relevant because larger sample sizes can lead statistically significant results for even negligible DIF. Therefore, assessing the magnitude of DIF is crucial to ascertain whether it is practically significant or not. To address this concern, various effect size measures have been developed and utilized to categorize DIF items in conjunction with statistical testing. Previous studies have demonstrated that incorporating effect size measures alongside statistical tests can help reduce Type I errors associated with large

sample sizes (G´omez-Benito, Dolores-Hidalgo, & Zumbo, 2013; Jodoin & Gierl, 2001; Roussos & Stout, 1996).

Various effect size measures have been developed and studied for DIF detection methods based on observed scores or IRT model. The observed score-based effect size measures include the delta metric used in Mantel-Haenszel (MH) DIF analysis method, standardized p-differences (STD P-DIF), and weighted-least-squares R^2 for logistic regression DIF (Dorans & Kulick, 1986; Holland & Thayer, 1988; Schmitt & Dorans, 1990; Shealy & Stout, 1993; Zumbo & Thomas, 1997). Based on the Mantel-Haenszel (MH) DIF analysis method, delta metric ($\Delta\alpha_{MH}$; Holland & Thayer; 1988) is one of the most widely used DIF effect size measure. The MH procedure uses observed score as the examinee matching criterion and provides both a significance test and estimate of the DIF magnitude. The purpose of the MH procedure is to compare the odds for success between reference and focal groups at a given level of the matching variable. The MH statistic has a chi-square distribution with one degree of freedom under the null hypothesis of no DIF and a significant MH_{χ^2} indicates DIF. Holland and Thayer (1988) proposed a measure of effect size using logarithmic transformation of the MH odds ratio. Using the delta metric, $\Delta\alpha_{MH}$ which is the logarithmic transformation of the odds ratio, Education Testing Service (ETS) classifies DIF items into 3 categories: A, B, and C (Zwick & Ercikan, 1989). More specifically, item labeled with A shows negligible DIF, where chi square is not significant and $|\Delta\alpha_{MH}| < 1$; a B item shows intermediate DIF, where chi square is significant and $1 < |\Delta\alpha_{MH}| < 1.5$, and a C item shows large DIF, where chi square is significant and $|\Delta\alpha_{MH}| \geq 1.5$.

Another measure, standardized p-differences (STD P-DIF), is proposed by Dorans & Holland (1993). The standardization approach compares the proportion of correct response for

the reference and focal group at each score level (Dorans & Holland, 1993; Dorans & Kullick, 1986). In the standardization approach, a weighting factor is used, and in most cases, it takes the value of the number of examinees in focal group. In order to detect DIF, the differences in proportions of success for focal group and reference group at each score level are weighted and then summed across the score levels. The formula for the STD P-DIF is

$$STD\ P-DIF = \sum_{m=1}^{M} w_m(P_{Fm} - P_{Rm}) / \sum_{m=1}^{M} w_m, \quad (8)$$

where w_m is the weighting factor at matching variable m, P_{Fm} and P_{Rm} are the proportions of correct response for the focal and reference group, respectively. The values of STD P-DIF can range from -1.0 to 1.0 and while positive STD P-DIF values indicate that items favor the focal group, negative STD P-DIF values indicate that items favor the reference group. Dorans and Holland (1993) proposed the following guidelines to interpret the DIF effect size: |STD P-DIF| < .05 indicates negligible DIF, .05 ≤|STD P-DIF| ≤ .10 indicates moderate DIF and may require further examination, |STD P-DIF| > .10 indicates large DIF and require a close examination.

Zumbo and Thomas (1997) proposed ΔR^2 as an effect size measure when logistic regression is used in DIF detection. They suggested the following guidelines to quantify the magnitude of DIF: ΔR^2< 1.3 indicates negligible DIF, .13 ≤ ΔR^2 ≤ .26 indicates moderate DIF, ΔR^2 ≥ .26 indicates large DIF. Later, Jodoin and Gierl (2001) proposed new threshold values for ΔR^2 based on the SIBTEST effect size measure ($\hat{\beta}$; Shealy & Stout, 1993) as follows: ΔR^2< .035 indicates negligible DIF, .035 ≤ ΔR^2 ≤ .070 indicates moderate DIF, ΔR^2 ≥ .070 indicates large DIF.

In addition to the observed score-based effect size measures, IRT model-based effect size measures based on the probability difference between the focal and reference groups have been proposed (Raju, 1988; Wainer, 1993). One measure to investigate DIF is Raju's area indices

(1988) that are obtained from the area between the ICCs for the reference and focal groups. When $P_R(\theta)$ and $P_F(\theta)$ indicate the probabilities of endorsing an item correctly for the reference and focal group, respectively, the signed and unsigned areas between the groups are defined as:

$$Signed\ Area = SA = \int_{-\infty}^{+\infty} [P_R(\theta) - P_F(\theta)]\, d\theta, \tag{9}$$

$$Unsigned\ Area = UA = \int_{-\infty}^{+\infty} |P_R(\theta) - P_F(\theta)|\, d\theta. \tag{10}$$

The SA can take positive or negative values and indicate uniform DIF where positive values indicate item favoring the reference group and negative values indicate item favoring the focal group. For nonuniform DIF items, the area between the ICCs is positive for some ability levels and negative for other ability levels. In this case, the total signed area might be small or even equal to zero because positive and negative DIF may cancel each other. Thus, the UA uses the absolute values of the differences between the ICCs and only takes positive values.

Wainer (1993) proposed four standardized DIF indices for the IRT models which is based on weighting group differences in probabilities by the proficiency distribution of the focal group. The four indices proposed by Wainer (1993) are defined as:

$$T(1) = \int_{-\infty}^{+\infty} [P_F(\theta) - P_R(\theta)]\, dG_F(\theta), \tag{11}$$

$$T(2) = N_F\, T(1), \tag{12}$$

$$T(3) = \int_{-\infty}^{+\infty} [P_F(\theta) - P_R(\theta)]^2\, dG_F(\theta), \tag{13}$$

$$T(4) = N_F\, T(3). \tag{14}$$

where $P_F(\theta)$ and $P_R(\theta)$ indicate the probabilities of endorsing an item correctly for the focal and reference group, respectively, $G_F(\theta)$ is the proficiency distribution for the focal group, and N_F is

the number of examinees in the focal group. The index of standardized impact, T(1), represent the average effect size for each examinee in the focal group. When multiple focal groups are included, T(2) represents a measure of total effect size for a certain focal group. In order to capture the magnitude of nonuniform DIF, T(3) represents a squared standardized effect size and T(4) is a total effect size. The thresholds for these standardized indices haven't been specified. These indices have been extended to analyze polytomous items (Kim et al., 2007) and to multidimensional tests in which the indices were evaluated in a multidimensional IRT (MIRT) model (Suh, 2016). Kim et al. (2007) suggested .05 cut off value for careful examination of DIF items which is based on the STD P-DIF.

The majority of DIF detection methods, both parametric and nonparametric, rely on the assumption of unidimensionality in the data. Accordingly, the matching variable is based on a unidimensional raw score or latent ability estimate. In the context of IRT, DIF occurs when the probability of answering an item correctly differs for examinees with the same latent ability or total score but belonging to different groups. However, in the DCMs, instead of providing a rank order on the latent continuum, latent attribute profiles are provided which indicate whether the examinee has mastered the measured attributes. Due to the differences between DCMs and traditional IRT models, the current DIF methods and effect size measures may overlook multidimensionality in the data, thus compromising the accuracy and comprehensiveness of DIF detection in DCMs. Therefore, DIF analysis specifically tailored for DCMs is necessary in order to address this limitation and enhance the effectiveness of DIF analysis in multidimensional contexts.

DIF Detection in DCM Framework

DIF analysis in DCMs can provide evidence that attribute-item interactions are invariant between groups (Hou et al., 2014). The presence of DIF items can lead to incorrect estimations of item parameters and latent attribute profiles, resulting in inaccurate interpretations (Hou et al., 2014; Paulsen et al., 2019). Therefore, it is necessary to conduct DIF analysis in order to establish item parameters or construct invariance. Several studies have investigated the detection of DIF in the DCM framework (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Liu et al., 2019; Ma, Terzi, & de la Torre, 2021; Svetina et al., 2018; Zhang, 2006).

In Zhang's (2006) study, the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) were examined as methods for detecting DIF under the DINA model. Two matching criteria, total scores and attribute mastery profiles were compared. The results indicated that using attribute profiles as the matching variable yielded better Type I error and power rates compared to using total scores. However, a limitation of this study was identified in that the presence of DIF items could introduce bias in the estimates of item parameters and attribute profiles. Consequently, utilizing these estimated attributes as matching criteria for the MH and SIBTEST methods may lead to biased DIF detection. Furthermore, it is important to note that this method is limited to detecting uniform DIF only.

Li (2008) conducted a study investigating DIF and differential attribute functioning (DAF) simultaneously using a modified higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004). DAF refers to the situation where the probability of mastering an attribute differs for examinees with the same attribute profile but from different groups. The findings indicated that the item parameter recovery was better than attribute parameter recovery.

Moreover, this method demonstrated better Type I error and power rates in comparison to the MH method.

Hou et al. (2014) investigated the effectiveness of the Wald test in detecting both uniform and nonuniform DIF in the DINA model. The researchers compared the Wald method with the MH and SIBTEST methods. The results demonstrated that the Wald method was an efficient method for DIF assessment and outperformed the MH and SIBTEST methods across various conditions. The study findings indicated that when the items had relatively high discrimination, the Type I error rates were at the nominal level for both small and large sample sizes. Regarding power, the DIF size and item discrimination significantly influenced the power rates. The Wald test exhibited medium to high power in detecting both uniform and nonuniform DIF when the DIF size was large, and the items were relatively discriminating. However, the Wald test showed inflated Type I errors and reduced power rates when the item discrimination was low, such as when the guessing and slipping parameters were as large as .3. Lastly, because of the separate calibrations performed for the Wald test, it was not affected by the percentage of DIF items in the test.

The LCDM-DIF method, proposed by Li and Wang (2015) aimed to address the multi-group variables by incorporating categorical, continuous, observed, or latent grouping variables into the analysis. This approach involved regressing item parameters on the grouping variables to detect DIF. In their simulations, Li and Wang compared the performance of the LCDM-DIF method with the Wald method introduced by Hou et al. (2014). The results indicated that the LCDM-DIF and Wald method demonstrated comparable performance in most conditions. However, the LCDM-DIF method outperformed the Wald method when the item discrimination was low.

Svetina et al. (2018) conducted a study to investigate the impact of Q-matrix misspecifications on the performance of three DIF methods (logistic regression, MH, and Wald method) in identifying DIF items within the DCMs using DINA model. They aimed to evaluate the performance of these methods when Q-matrix misspecification occurs in one group. The findings of the study indicated that logistic regression and MH methods demonstrated better control of Type I error compared to the Wald test. However, in terms of power, both logistic regression and Wald methods showed higher rates. Moreover, the results indicated that the performance of the methods was influenced by the complexity of the Q-matrix and item structures, leading to varying outcomes for different methods.

In their study, Liu et al. (2019) investigated the effectiveness of the Wald method in detecting DIF using different types of covariance matrices. They compared the performance of the method when the covariance matrix estimated using complete information approach and that of the item-wise information matrix. The results indicated that the covariance matrix estimated using a complete information approach yielded better Type I error rates and power rates compared to the item-wise information matrix. This suggested that considering the complete information approach for estimating the covariance matrix can enhance the accuracy of DIF detection using Wald test.

Finally, Ma et al. (2021) introduced a multiple-group generalized deterministic inputs, noisy “and” gate (MG-GDINA) model and evaluated the performance of the likelihood ratio (LR) test and Wald test in detecting DIF under MG-GDINA model. The findings revealed that the Type I error rates were well controlled when the items demonstrated moderate to high quality. However, the power to detect DIF was generally low across most conditions, except in

favorable scenarios characterized by large DIF magnitude, large sample sizes, high item quality, and the presence of uniform DIF.

Overall, DCM model-based methods such as modified HO-DINA, Wald test, and LCDM-DIF methods exhibited notable advantages over traditional method such as the MH and SIBTEST. These model-based approaches consistently demonstrated more effective control of Type I error and higher power rates, making them superior choices for detecting DIF in DCMs. These model-based methods identify nonuniform DIF, thus enabling a more comprehensive assessment of the DIF across distinct groups. Furthermore, DIF studies also showed that the effectiveness of the Wald method relies on the item discrimination (Hou et al., 2014; Li & Wang, 2015; Liu et al., 2019; Ma et al., 2021) and the performance of the Wald method when items are not discriminating might be improved by employing a more complex variance-covariance matrix (Liu et al., 2019).

Effect Size Measures in DCM Framework

In the DCM framework, DIF detection methods have been developed based on the significant test (Hou et al., 2014; Li & Wang, 2014) and there have only been two studies that adopted effect size measures to detect DIF (Feng, 2021; George & Robitzsch, 2014). George and Robitzsch (2014) adopted an effect size measure based on the unsigned area (UA) originally introduced by Raju (1990). This effect size measure is based on the difference of the item response functions between groups:

$$UA_j = \sum_{l=1}^{L} w(\alpha_l) \left| P\left(X_j = 1 | \boldsymbol{\alpha}_l, g_1\right) - P\left(X_j = 1 | \boldsymbol{\alpha}_l, g_2\right) \right|, \quad (15)$$

where $w(\boldsymbol{\alpha}_l) = \frac{1}{2}\left(P(\boldsymbol{\alpha}_l | G = g_1) + P(\boldsymbol{\alpha}_l | G = g_2)\right)$, and $P(\boldsymbol{\alpha}_l | G = g_1)$, $P(\boldsymbol{\alpha}_l | G = g_2)$ are the probability of being in latent class l given group 1 and group 2, respectively. They adopted the

cut-off values of .059 to distinguish between negligible and moderate DIF, and .088 for moderate and large DIF, as suggested by Jodoin and Gierl (2001) in the context of the three parameter IRT model. They conducted the DIF analysis using a large-scale assessment in order to assess the item parameter invariance across different groups before applying multiple group DCMs. However, the applicability and effectiveness of the adopted cut-off values, which are derived from the IRT framework, necessitates further investigation to be used for effect size measures based on multidimensional attribute profiles.

Feng (2021) proposed an effect size measure that is defined as the weighted between-group difference in the success probability on an item. While UA_j effect size (George and Robitzsch, 2014) weights over the distribution of the attribute pattern for both focal and reference group, the proposed effect size measure uses focal group distribution only that is typically disadvantaged group and is the subject of the majority of DIF studies. It defined as:

$$SPD_j = \sum_{l=1}^{L} \left[P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_F - P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_R \right] * \ P(\boldsymbol{\alpha}_l|G = F) \tag{16}$$

where SPD_j indicates the signed probability difference for item j; $P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_F$ and $P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_R$ are the probability of answering item j correctly for examinees having attribute profile $\boldsymbol{\alpha}_l$ in the focal group and reference group, respectively. The weighting factor, $P(\boldsymbol{\alpha}_l|G = F)$, indicates the probability of being in latent class l in the focal group. While positive values indicate the item favors the focal group, negative values indicate the item favors the reference group. This also allows the cancelation of the DIF effect for non-uniform DIF so that it is possible to have an SPD of zero or even a very small one. Thus, based on the absolute values of the probability differences, an unsigned measure is also proposed along with a root mean square probability difference:

$$UPD_j - abs = \sum_{l=1}^{L} \left| P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_F - P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_R \right| * \ P(\boldsymbol{\alpha}_l|G = F) \quad (17)$$

$$UPD_j - sqrt = \sqrt{\sum_{l=1}^{L} \left(P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_F - P\left(X_j = 1|\boldsymbol{\alpha}_l\right)_R \right)^2 * \ P(\boldsymbol{\alpha}_l|G = F)}. \quad (18)$$

Feng (2021) examined the estimation accuracy of SPD and UPD under various conditions with the DINA model and found that the accuracy of SPD and UPD estimates is satisfactory. They also used these indices combined with the Wald test in order to examine the impact of including effect size measures on Type I and power rates of DIF detection. The results showed that including effect size measures could substantially reduce the Type I error rates and led to a reduction in power for detecting smaller DIF, while the power rates of large DIF were not affected. As the SPD and UPD indices are based on between-group differences in probabilities, the STD P-DIF threshold values of .05 and .10 were adopted to classify DIF items into negligible, moderate, and large DIF categories (Dorans and Holland, 1993) and the adequacy of the existing classification schemes for the standardized P-difference and MH effect size were evaluated (Feng, 2021). The results showed that both MH effect size measure and the SPD/UPD indices were able to differentiate three levels of DIF magnitude in the presence of uniform DIF. Regarding non-uniform DIF, the MH effect size measure failed to detect the non-uniform DIF and the UPD indices worked well as those are an unsigned index, which prevents the cancelation of the DIF effect.

The studies have demonstrated the effectiveness of suggested effect size measures in distinguishing different categories of DIF by utilizing threshold values based on group probability differences. However, further investigation is needed to understand the impact of DIF items on the primary purpose of DCMs, which is examinee classification. This research would

provide insights into how various scenarios related to DIF influence classification accuracy and aid in determining threshold values for identifying DIF items. Moreover, the current studies primarily focused on the DINA model which is a noncompensatory model requiring mastery of all attributes for item endorsement. Considering that general DCMs are more complex with additional parameters to estimate, it is crucial to assess the performance of effect size measures within a broader DCM framework and explore criteria for determining the practical significance of DIF.

This dissertation has two goals. In Study 1, we investigate the performance of Wald DIF detection using LCDM as a general framework under a variety of conditions. The Wald DIF detection method was proposed by Hou et al. (2014) to determine if the item parameters were equal between the groups. In the proposed procedure, separate calibrations are performed for reference and focal groups, which avoids contamination from DIF items, so test purification is not necessary (Hou et al., 2014). We will set up simulation conditions that aim to cover some critical factors of DIF studies to examine the performance of Wald DIF in the LCDM model. Then, we will compare the performance of the Wald method and the likelihood ratio test (LR) in detection of DIF based on the LCDM model.

As the second aim of this dissertation, in Study 2, we will investigate the criteria for identifying the degrees to which DIF is practically significant. We examine thresholds based on the degree to which DIF items impact DCM classifications and reliability in order to provide practitioners with criteria to flag items. Specifically, we aim to define the big, medium, and small impact of DIF in terms of classifications. We will compare the classification accuracy rates with the effect size measures developed for DCMs (e.g., UA_j) in order to define the magnitude of DIF in terms of classifications.

Overview of Chapters

This dissertation is organized as follows. Chapter 1 provides an introduction and literature review. Chapter 2 and 3 are dedicated to standalone studies focusing on DIF analysis in DCMs. In Chapter 2, the performance of the Wald method is evaluated for DIF detection within the framework of LCDM. Chapter 3 focuses on effect size measures in DCMs by investigating DIF effect on DCM classifications. Chapter 4 summarizes and discusses the findings from both studies and concludes with an overview of future study directions.

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.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Lawrence Erlbaum Associates.

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied psychological measurement*, *17*(4), 335-350.

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement* (p. 328). Taylor & Francis.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 50–87). Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement*, *23*(4), 355-368.

Feng, Y. (2021). *Effect Size Measures for Differential Item Functioning in Cognitive Diagnostic Models*. Indiana University.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405.

Gómez-Benito, J., Hidalgo, M. D., & Zumbo, B. D. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items. *Educational and Psychological Measurement*, *73*(5), 875-897.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313-334.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74(2)*, 191–210.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure In Wainer H & Braun HI (Eds.), Test validity (pp. 129–145). Hillsdale, NJ, US.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, *14*(4), 329-349.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of educational measurement*, *44*(2), 93-116.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement, 56*(5), 746-759.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning*. Doctoral dissertation, University of Georgia, Athens, GA.

Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28-54.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, *10*, 1137.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting Differential Item Functioning Using Multiple-Group Cognitive Diagnosis Models. *Applied Psychological Measurement*, *45*(1), 37-53.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748.

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267-281.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error Performance. *Journal of Educational measurement*, *33*(2), 215-230.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6:4, 219-262.

Rupp, A., Templin, J. and Henson, R. A. (2010) Diagnostic Measurement: Theory, Methods, and Applications. Guilford Press.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67-81.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement*, *53*(4), 403-430.

Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika*, *44*, 313-349.

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in psychology*, *9*, 696.

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

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic monitoring of skill and knowledge acquisition*, 453-488.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251-275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 123–135). Hillsdale, NJ: Erlbaum.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model*. Doctoral dissertation, The University of North Carolina at Greensboro.

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.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of educational measurement*, *26*(1), 55-66.

CHAPTER 2

AN INVESTIGATION OF DIFFERENTIAL ITEM FUNCTIONING IN DIAGNOSTIC CLASSIFICATION MODELS [1]

[1] Zor, S. and M.J. Madison. To be submitted to *Journal of Educational Measurement*.

Abstract

Differential item functioning (DIF) analysis is conducted to ensure test fairness and validity. Existing diagnostic classification models (DCMs) based DIF assessment methods have been developed to consider the multidimensional and binary nature of the attributes when examining DIF. This study focuses on extending the Wald DIF detection method to a general DCM framework, LCDM, and evaluating its performance through a simulation study. Additionally, the performance of the Wald method was compared with the LRT method. Results demonstrated that Wald method effectively controls Type I errors across various conditions, outperforming the LRT method, especially under small unequal sample sizes and short tests. Both methods showed better Type I error rates with equal sample sizes and base rates. While the Wald and LRT methods effectively detected DIF items with large DIF magnitude, power rates were relatively low for smaller DIF magnitudes. The findings also revealed the effect of DIF type on power rates, with higher power observed for conditions involving intercept only DIF or all parameters DIF (uniform DIF) compared to main effect and interactions DIF (nonuniform DIF).

Introduction

Diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010), also known as cognitive diagnosis models, analyze the multidimensional content of assessments to provide fine-grained categorical information (mastery/non-mastery) about examinees' knowledge components, called *attributes* (Rupp & Templin, 2008). To this point, DCMs have been proposed as viable models for practice (e.g., Bradshaw et al., 2014; de la Torre & Douglas 2008; Junker & Sijtsma, 2001; Madison & Bradshaw, 2018; Rupp, Templin, & Henson, 2010); however, compared to methodological developments, relatively few practical applications of these models exist. Lack of practical implementation of DCMs may be attributed to limited research addressing key psychometric questions required to establish validity and fairness, including differential item functioning (DIF) methods.

In the context of DCMs, DIF occurs when the probability of answering an item correctly differs for examinees with the same pattern of attribute mastery, or same latent *attribute profile*, but from different groups (Hou, de la Torre, & Nandakumar, 2014). This reflects the potential test bias with items functioning unequally across groups after controlling for attribute mastery and becomes a threat to test fairness. The existence of DIF items in DCMs could result in inaccurate estimates of item parameters and attribute profiles that render the interpretations based on examinees' latent profiles inaccurate (Hou et al., 2014). DIF analyses assess item-attribute invariance across different groups, and detecting and eliminating DIF items is a vital step of diagnostic assessment construction.

Currently, limited research has focused on DIF detection in the DCM framework (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Liu et al., 2019; Ma, Terzi, & de la Torre, 2021; Zhang, 2006). Zhang (2006) explored the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and the

simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) methods using different matching criteria to detect DIF under the DINA model. It was found that using attribute profiles as a matching variable resulted in better detection rates than using total scores. However, one of the limitations of this study is that the presence of DIF items in a test may affect the item parameter and attribute profile estimates, used as matching variable, which in return could affect DIF detection. Li (2008) examined DIF and differential attribute functioning (DAF) simultaneously using a modified higher-order DINA model (de la Torre &Douglas, 2004). Results indicated better item parameter recovery and improved detection rates compared to the MH method. Hou et. al. (2014) investigated the Wald method for detecting uniform and nonuniform DIF in the DINA model. The Wald method outperformed the MH and SIBTEST methods, demonstrating efficient DIF assessment. However, the Wald method showed inflated Type I errors and reduced power when items were low discriminating.

DCM-based methods have been found to effectively detect DIF. However, most of these methods have been based on the DINA model which posits one specific theory about attributes in order to simplify the DCM estimation. Applications have demonstrated educational assessment data are not expected to always, or even often, follow this simplified structure, so it is necessary to assess the performance of the methods under a more general DCM. There are two studies that have used general DCMs to investigate DIF. Li and Wang (2015) assessed DIF by regressing item parameters on grouping variables under the log-linear DCMs (LCDM; Henson, Templin, & Wilse, 2009) estimated through Markov chain Monte Carlo (MCMC). Ma et al. (2021) developed a multiple-group generalized deterministic inputs, noisy “and” gate (MG-GDINA) model and compared the performance of the likelihood ratio (LR) test and Wald test in detecting DIF under MG-GDINA model. The study showed that the Type I error rates were

relatively well controlled when the items are moderate to high quality and the power in detecting DIF was low in most of the conditions except the favorable conditions where large DIF magnitude and sample size, high item quality and uniform DIF exist. One of the limitations of this study is that there were some critical factors to DIF studies that were fixed such as assumed only equal sample sizes and distributions for groups, fixed test length and number of attributes. Also, the MG-DINA model is based on the identity link function and there are other DCMs based on alternative link functions (i.e., logit, log). Therefore, further research is needed to generalize the findings with extended simulation conditions and models of other link functions.

Thus, in this study, our purpose is to investigate the performance of the Wald DIF detection method after extending it to the general framework, namely LCDM, under a wide range of conditions. Then, we will compare the performance of the Wald method and likelihood ratio test (LRT) in detection of DIF based on the LCDM model.

Method

Log-linear Cognitive Diagnostic Model

In this study, we used LCDM that provides a general framework to model the relationship between item responses and attributes. A key feature of the model is that core DCMs can be represented by placing constrains on the LCDM parameters, which makes these models nested within the LCDM. In the LCDM framework, the probability of a correct response to item *i* is conditional on examinee *e*'s attribute profile α_e. Let X_{ei} be a response of an examinee *e* to item *i*, the general form of LCDM item response function is written as:

$$P(X_{ei} = 1|\boldsymbol{\alpha_e}) = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda_i^T h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i})\right)}{1 + \exp\left(\lambda_{i,0} + \boldsymbol{\lambda_i^T h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i})\right)} \quad (1)$$

where $\lambda_{i,0}$ is intercept and $\boldsymbol{\lambda_i^T}$ is a vector of size $(2^A - 1) \times 1$ with all main effects and possible interaction parameters for item *i*; *A* is the number of attributes; $\boldsymbol{q_i}$ represent the Q-matrix entries

for item *i*, and $\boldsymbol{h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i})$ is a vector of size $(2^A - 1) \times 1$ representing a set of linear combinations of $\boldsymbol{\alpha_e}$ and $\boldsymbol{q_i}$. $\boldsymbol{\lambda_i^T h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i})$ is written as:

$$\boldsymbol{\lambda_i^T h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i}) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ea}\, q_{ia} + \sum_{a=1}^{A-1} \sum_{b=a+1}^{A} \lambda_{i,2,(a,b)} \alpha_{ea} \alpha_{eb} q_{ia} q_{ib} + \cdots \tag{2}$$

where $\lambda_{i,1,(a)}$ represents the main effect of Attribute *a* on the item *i*, and $\lambda_{i,2,(a,b)}$ represents the two-way interaction effect of Attribute *a* and Attribute *b*. The right side of the equation includes all main effect parameters and all possible interaction parameters. Item parameters are present if the linear combination of $\boldsymbol{\alpha_e}$ and $\boldsymbol{q_i}$ equals 1. An intercept, $\lambda_{i,0}$, represents the log-odds of a correct response probability for examinees who have not mastered any of the attributes measured by item *i*, and the main effect, $\lambda_{i,1,(a)}$, represents the increase in log-odds of a correct response probability for examinees who have mastered the attribute *a* measured by item *i*. Constraints are placed on the main effect parameters so that examinees' correct response probabilities increase as they master an attribute, i.e., $\lambda_{i,1,(a)} > 0$.

Wald Method for DIF Detection

The Wald test is used by Hou et al. (2014) to detect DIF in the DINA model. To implement the Wald test, item parameters of the model are estimated separately for the focal and reference groups. Then, the null hypothesis that the item parameters of the focal and reference groups are equal is tested. In the LCDM model, the item parameter estimates for item i and the variance-covariance matrix of the item parameters are represented as follows:

$$\hat{\mathbf{v}}_i = (\hat{\mathbf{v}}_{Ri}, \hat{\mathbf{v}}_{Fi})' = \left(\left(\hat{\lambda}_{i,0}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{i}} \right)_R, \left(\hat{\lambda}_{i,0}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{i}} \right)_F \right)' \tag{3}$$

$$Var(\hat{\mathbf{v}}_i) = \begin{pmatrix} Var(\hat{\mathbf{v}}_{\mathrm{R}i}) & 0 \\ 0 & Var(\hat{\mathbf{v}}_{\mathrm{F}i}) \end{pmatrix} \tag{4}$$

The Wald statistic to test the equality of item parameters of the groups is computed as:

$$W_i = [\mathbf{R}_i\mathbf{v}_i]'\{\mathbf{R}_i\, Var(\mathbf{v}_i)\mathbf{R}_i'\}^{-1}[\mathbf{R}_i\, \mathbf{v}_i] \quad (5)$$

where $\mathbf{R}_i$ is a contrast matrix. For each group, $2^{\boldsymbol{A_i}}$ item parameters are estimated, and the dimension of the contrast matrix $\mathbf{R}_i$ is $2^{\boldsymbol{A_i}} \times 2^{\boldsymbol{A_i+1}}$, A is the number of attributes measured by item *i*. For example, if an item measures two attributes (A_i=2), the contrast matrix of $\mathbf{R}_i$ is:

$$\mathbf{R}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \quad (6)$$

Under the null hypothesis of H_0: $\mathbf{R}_i\mathbf{v_i} = \mathbf{0}$, W_i is asymptotically distributed as a chi-square distribution with $2^{\boldsymbol{A_i}}$ degrees of freedom.

Likelihood Ratio Test for DIF detection

The likelihood ratio test is an IRT model-based method that has been used to detect DIF (LRT; Thissen, Steinberg, & Wainer, 1993). This method compares the two nested IRT models in which the compact model constraints the item parameters for all items to be the same across reference and focal groups, and augmented model constraints all the item parameters except the studied item to be the same in both groups. The remaining items in the augmented model which constrained to be equal for parameter estimates for both groups serve as anchor items. The likelihood ratio statistic can be written as:

$$G^2 = (-2lnL_C) - (-2lnL_A) \quad (7)$$

here L_C is the loglikelihood of the compact model, and L_A is the loglikelihood of the augmented model. The test statistic follows a chi square (χ^2) distribution with the degrees of freedom that equal to the difference in the number of parameters estimated in each model.

Simulation Study

The simulation study was designed to examine the performance of the Wald DIF detection method under various conditions. It aims to provide insights to researchers and test developers about conditions required for the Wald DIF method to effectively detect DIF items under the LCDM model.

We manipulated key factors including base rates, sample sizes, test length, percentage of DIF items, DIF magnitude, and DIF type. We fixed the number of attributes at 4 across the simulation conditions. The attribute tetrachoric correlations were fixed at .50, and the reference group correct response probabilities for masters (ranges between .70 to .80), non-masters (ranges between .20 to .30), partial masters (ranges between .50 to .60) to be the same across conditions. We selected those item response probability ranges to have a test with moderate difficulty and with moderately discriminating items. The generating and estimation model was LCDM. Table 2.1 summarizes the manipulated factors for this study and these simulation conditions are described in detail below.

Manipulated Factors

Base Rate

Ability distribution differences between the focal and reference groups are shown to be an important factor in DIF detection studies (Li, 2008; Mazor, Clauser, & Hambleton, 1992; Narayanon & Swaminathan, 1996; Paulsen et al., 2020). We manipulated the base-rates to have groups with different underlying attribute distributions. As the base-rate of an attribute is the proportion of examinees who have mastered the attribute, groups with differing base-rates can enable us to study the impact of unequal examinee distributions in DIF analysis. We simulated

the reference and focal groups having balanced mastery proportions (.50, .50) and unbalanced mastery proportions (.75, .25) where reference group has the high base-rate of mastery.

Sample Size

We manipulated sample sizes to examine the extent to which equal and unequal sample sizes of the reference and focal groups influenced the DIF analysis. We used two levels of equal sample sizes for each group ranging from small to sufficient (500/500 and 1000/1000) and two levels of unequal sample sizes for reference and focal group (750/250 and 1500/500), respectively. Under the unequal sample sizes, the sample sizes of reference group are larger than focal group sample sizes.

Test Length

We manipulated the test length and used two levels: 24 items and 48 items test. In the short test condition, a balanced Q-matrix was used to measure each attribute with eight items (4 simple items and 4 complex items measuring two attributes) as shown in Table 2.2. We doubled the number of items for the longer test condition (48 items) and each attribute is measured with 16 items that consists of 8 simple items and 8 complex items measuring two attributes.

Percentage of DIF items

We manipulated the percentage of DIF items in the test and used four levels: 0, 12.5%, 25%, and 50%. These percentages fall within the range of previous DIF studies in DCM framework (Hou et al., 2014; Li & Wang, 2015; Liu et al., 2019; Ma, Terzi, & de la Torre, 2021; Paulsen et al., 2020). In order to simulate DIF for the short test condition, we selected 3 items (2 simple, 1 complex items) for 12.5% DIF percentage, 6 items (3 simple, 3 complex) for 25% DIF percentage and 12 items (6 simple, 6 complex) for 50% DIF percentage. We doubled the number of items for the long test condition and selected 6 items (4 simple, 2 complex), 12 items (6

simple, 6 complex) and 24 (12 simple, 12 complex) for 12.5%, 25%, and 50% DIF conditions, respectively. Then, we fixed these items over conditions, and used the same items as DIF items.

DIF Magnitude

The differences in the LCDM item parameters ($\Delta_{\lambda_{i,0}}, \Delta_{\boldsymbol{\lambda_i}}$) between the focal and reference groups is defined as the DIF magnitude. $\Delta_{\lambda_{i,0}}$indicates the difference in the intercept parameters between the focal group and reference group, and $\Delta_{\boldsymbol{\lambda_i}}$ indicates the difference in the main effects and possible interaction terms for item *i* between the focal and reference groups (i.e., $\Delta_{\lambda_{i,0}} = (\lambda_{i,0})_F - (\lambda_{i,0})_R$ and $\Delta_{\boldsymbol{\lambda_i}} = (\boldsymbol{\lambda_i})_F - (\boldsymbol{\lambda_i})_R$).Most of the previous studies used .05 and .10 as small and large DIF (Zhang, 2006; Hou et al., 2014; Li & Wang, 2015; Svetina et al., 2018), while some of the others also examined additional levels of .15 and .20 (Paulsen et al., 2020, Feng 2021). Three levels of DIF magnitude are used in this study: small (.3), medium (.5), and large (.8). Note that DIF sizes used in the previous studies reflected the probability differences because the item parameters were guessing and slipping parameters. The DIF sizes used in this study are corresponding values in the logit scale. The focal group's item parameters were obtained by adding and subtracting DIF size from the reference group's item parameters depending on the design.

DIF Type

We manipulated the DIF types based on the LCDM item parameters where we introduced DIF to only intercept, only main effects and interactions, and all parameters. For intercept only conditions, focal group's intercept parameters were obtained by subtracting DIF magnitudes from reference group's intercept parameters, and other item parameters didn't change ($\Delta_{\lambda_{i,0}} < 0, \Delta_{\boldsymbol{\lambda_i}} = 0$). For main effect and interactions conditions, the probabilities for nonmasters didn't change between groups and the probabilities for masters in focal group were decreased ($\Delta_{\lambda_{i,0}} =$

$0, \Delta_{\boldsymbol{\lambda_i}} < 0$). Lastly, we manipulated all item parameters by decreasing the focal group's item parameters relative to reference group item parameters ($\Delta_{\lambda_{i,0}} < 0, \Delta_{\boldsymbol{\lambda_i}} < 0$). This condition corresponds to manipulating all item parameters by having lower intercept, main effects and interaction parameters (for complex items) in the focal group. These conditions were selected to reflect the situations where reference group is favored by items (DIF against to focal group). Table 2.3 summarizes the combination of DIF magnitude and DIF types for this study.

Overall, six key factors were manipulated: base-rates (balanced and unbalanced mastery proportions), sample sizes (equal and unequal sample sizes with two levels for focal and reference groups), test length (short and long), percentage of DIF items (12.5%, 25%, 50%), DIF magnitude (small, medium, large), and DIF type (intercept only, main effects and interactions only, and all parameters). Crossing these 6 factors yielded 432 simulation conditions. For no-DIF conditions, crossing 3 factors (base rates, sample size and test length) yielded 16 conditions, total of 448 simulation conditions. Under each condition, 100 replications were generated. The simulation study was implemented in R (R Core Team), and the *GDINA R* package (Ma & de la Torre, 2020) was used to perform DIF analysis. Hou et al. (2014) calculated the covariance matrix for the Wald statistic based on the item-wise information matrix and ignored the structural parameters. However, the Type I error rates were inflated under certain conditions such as small sample size and low discriminating item conditions because of the underestimated item parameter covariance matrix (Hou et al. 2014). Liu et al. (2019) showed that using an outer-product of gradient (OPG; Philipp et al., 2018) method when all parameters are used to calculate the covariance matrix resulted in better Type I error rates for the Wald DIF detection method based on the DINA model. Therefore, this study considers OPG method for the variance-covariance matrix calculation.

Evaluation Criteria

Two criteria were used to evaluate the performance of the Wald method and LRT in detecting DIF items based on the general DCMs: Type I error and power rates.

Type I Error. Type I error indicates that an item is flagged as having DIF when it is a DIF-free item. Type-I error rates were calculated as the proportion of times a DIF-free item incorrectly flagged as DIF item out of the 100 replications. Then, the Type I error rates were averaged across all the DIF-free items for each condition. Based on the exact binominal distribution, the Type I error rates would be expected to fall within $p \pm 1.96\,(p(1-p)/n)^{1/2}$ if the test were adhering well to the nominal level of p with *n* replications. In our study with 100 replications, the Type I error rates are expected to fall between .007 and .093 with 95% chance at the .05 alpha level.

Power. Power rates in the DIF context were defined as the proportion of times a DIF item correctly flagged as having DIF out of 100 replications. The power rates of .80 or above were used to indicate excellent power (Cohen, 1992; Hou et al., 2014).

Simulation Study Results

We present the results of Type I error and power rates for Wald and LRT methods under 448 simulation conditions including: base rates for reference and focal group (2 levels: equal and unequal), sample sizes (4 levels: equal, 500-500, 1000-1000, and unequal, 750-250, 1500-500 reference group's samples size is larger), test length (short and long), percentage of DIF items (0, 12.5%, 25%, 50%), DIF magnitude (small, medium, and large: 0.3, 0.5, 0.8), and DIF type (intercept only, main effects and interactions, and all parameters). Results were summarized based on 100 replications for each condition. The Type I error and power rates results were

investigated to assess the effectiveness of Wald test in LCDM model for detecting DIF. The performance of Wald and LRT methods for DIF detection was compared.

Type I Error Rates

Table 2.4 shows the Type I error rates for the Wald and LRT methods under the no DIF conditions across the test lengths, sample sizes and base rates of the reference and focal groups. Under the short test conditions, the Type I error rates in no-DIF conditions were well controlled for both methods, ranging between 0.038 and 0.054 for Wald method, and between 0.048 and 0.08 for LRT method. The Type I error rates for LRT method were slightly higher than the nominal level of 0.05. Under the long test conditions, the Type I error rates for Wald test were lower than the nominal level, ranging between 0.004 and 0.026, and had deflated Type I error rates when reference and focal groups had unequal and small sample sizes ($N_R = 750$, $N_F = 250$). The Type I error rates for LRT were well controlled, ranging between 0.049 and 0.059. As the sample size increased, the Type I error rates got closer to the nominal level for both methods. Under the unequal base rate conditions, the methods had good performance in controlling Type I error rate, except for the Wald method in the small and unequal sample size (0.004).

Table 2.5 and Table 2.6 present the Type I error rates for the Wald and LRT methods under the equal base rate and short test conditions across the various DIF percentages, DIF magnitudes, and the DIF types for the equal and unequal sample sizes of the reference and focal groups, respectively. The Type I error rates were close to the nominal level for both Wald method (between 0.034 and 0.064) and LRT (between 0.043 to 0.073) under the equal base rate and equal sample sizes. Furthermore, Type I error rates were better controlled under the equal sample sizes compared to the unequal sample sizes, where the Type I error rates for Wald method was between 0.035 and 0.078, and those for LRT method was between 0.045 and 0.09

when groups have unequal sample sizes. Specifically, under the small sample sizes for both equal and unequal conditions (i.e., N_R =500, N_F =500 or N_R =750, N_F =250), when 50% of DIF items introduced to all item parameters with 0.8 DIF magnitude, both methods got the most inflated Type I errors.

Table 2.7 and Table 2.8 present the Type I error rates for the Wald and LRT methods under the unequal base rate and short test conditions across the various DIF percentages, DIF magnitudes, and the DIF types for the equal and unequal sample sizes of the reference and focal groups, respectively. Overall, both methods performed better in the equal base rate conditions compared to unequal base rate conditions (see Table 2.5 and Table 2.6). The Wald method under the unequal base rate conditions performed well in controlling Type I error rates, ranging between 0.038 and 0.081. The highest Type I error observed in the condition where 50% of DIF items with 0.8 DIF magnitude introduced to main effect and interaction parameters under the unequal large sample sizes (0.081). On the other hand, the Type I error rates were inflated for the LRT method under the small and unequal sample sizes (N_R =750, N_F =250) especially when 50% of DIF items present, ranging from 0.094 to 0.113. Under other sample sizes of the DIF conditions, the Type I error rates of LRT were well controlled and all higher than the nominal level.

For the long tests, Table 2.9 and Table 2.10 present the Type I error rates for the Wald and LRT methods under the equal base rates across the various DIF conditions for the equal and unequal sample sizes, respectively. The Wald method performed well in controlling Type I error rates for long tests under the equal base rates across the conditions, except the unequal and small sample sizes with deflated Type I error rates ranged between 0.00 and 0.006 (see Table 2.10). The Type I error rates of the Wald method were lower than the nominal level across all

conditions. The Type I error rates for the LRT method were well controlled and close to the nominal level across all conditions, ranging between 0.04 to 0.065 under the equal base rates.

Table 2.11 and Table 2.12 present the Type I error rates for the Wald and LRT methods under the unequal base rates and long tests across the various DIF conditions for the equal and unequal sample sizes, respectively. Similar results were observed for the unequal base rate conditions under the long tests, however, both methods performed slightly better in the equal base rate conditions compared to unequal base rate conditions. The Type I error rates ranged between 0.001 and 0.023 for Wald method, and between 0.045 and 0.07 for LRT method. Similarly, under the unequal and small sample size conditions, the Wald method had deflated Type I error rates, ranging between 0.001 and 0.004.

Power Rates

Table 2.13 and Table 2.14 present the power rates for the Wald and LRT methods under the equal base rate conditions across the various DIF percentages, DIF magnitudes, and the DIF types for the equal and unequal sample sizes of the reference and focal groups, respectively. Figure 2.1 and Figure 2.2 are also presented in order to visualize the results for both methods. From the figures, it is clear that the sample sizes influenced the power rates, as the sample sizes increased the power rates of detecting DIF for both methods increased. For example, when groups have equal sample sizes, the average power rates of Wald method were 0.64 for intercept only DIF, 0.44 for main effect and interaction DIF, and 0.64 for all parameter DIF when groups sample sizes are 500, and that of 0.83, 0.67 and 0.81 when groups sample sizes are 1000. The trend was similar under the unequal sample sizes; however, the power rates were higher for the conditions where focal and reference groups having equal sample sizes compared to groups having unequal sample sizes.

The power rates were also impacted by the DIF magnitude. The power rates increased as the DIF magnitude increased across all conditions for both Wald and LRT method. The results showed that the power rates when DIF magnitude was 0.8 were more than twice of those when DIF magnitude was 0.3 across all conditions. Both methods had excellent power, over 0.80, when DIF magnitude was 0.8 and DIF was introduced to intercept only and all parameters. Under these two DIF type conditions, the methods also performed well (over 0.80) when DIF magnitude was 0.5 and sample sizes are large ($i.e., N_R$ =1000, N_F =1000 or N_R =1500, N_F =500). When DIF magnitude was 0.3, the power rates were between 0.08 and 0.61 for Wald method, and 0.12 and 0.63 for LRT method. The power rates increased for the DIF magnitude of 0.5, ranged from 0.23 to 0.98 for Wald method, and 0.34 to 0.98 for LRT method.

The DIF type also had some effect on the power rates of the methods. The power rates were higher when DIF was introduced to intercept only conditions or all parameter conditions compared to main effect and interaction DIF type conditions. The average power rates of Wald method were 0.69 for intercept only conditions, 0.68 for all parameter conditions and 0.50 for main effect and interactions condition. The power rates of LRT method were 0.68, 0.66 and 0.51 for corresponding DIF type conditions, respectively. Lastly, the DIF percentages didn't affect the power rates of both methods.

Table 2.15 and Table 2.16 present the power rates for the Wald and LRT methods under the unequal base rate conditions across the DIF type and sample size conditions. The power rates under the unequal base rate conditions followed similar trends with the equal base rate conditions regarding the effect of sample size, DIF magnitude and DIF type. However, when we focus on the impact of the base rates, the power rates were slightly lower for the groups having unequal base-rates compared to groups having equal base-rates. In comparison to equal base rates with

unequal base rates, the power rates of Wald method decreased by 20% for intercept only DIF, 20% for main effect and interaction DIF, and 15% for all parameter DIF. Similarly, the power rates of LRT method decreased by 11% for intercept only DIF, 13% for main effect and interaction DIF, and 8% for all parameter DIF when compared to equal and unequal base rates. Lastly, the power rates followed similar trends across the conditions when a longer test was used, and slightly higher power rates were observed for methods in detecting DIF. Related Tables 2.17, 2.18, 2.19 and 2.20 and Figures 2.3 and 2.4 are given to summarize the power rates of both methods under long test conditions.

Discussion

DCM based DIF assessment methods have been developed to consider the multidimensional and binary nature of the attributes when investigating DIF. These methods have shown effectiveness in detecting DIF; however, most of these methods rely on the DINA model, which simplifies DCM estimation by assuming a specific theory about attributes. Real-world educational assessment data often deviates from this simplified structure, highlighting the need to evaluate DCM-based DIF methods under a more general framework. In this study, we assess the performance of the Wald DIF detection method after extending it to the LCDM framework. We conducted a simulation study to systematically investigate DIF in a general DCM model using Wald method and compared the performance of it with LRT method.

Results from our study indicated that the Wald method effectively controlled Type I errors across different conditions under the short tests and had deflated Type I errors for the long tests when groups had small and unequal sample sizes. The LRT method, however, exhibited inflated Type I errors in short length conditions where the groups had small and unequal sample sizes and performed better under the long test conditions. The Wald test tended to be more

conservative, with Type I error rates below the nominal level, and the LRT method tended to be more liberal, with the Type I error rates above the nominal level for most of the conditions. Both methods showed better Type I error rates under equal sample size and base rates. Overall, the Wald method had better performance in controlling Type I errors under short tests even under small unequal sample sizes, which makes it a reasonable choice.

The results also showed that the Wald and LRT methods performed well in detecting DIF items when the DIF magnitude was large (i.e., 0.8). However, power rates were relatively low for both methods in conditions with smaller DIF magnitudes. Overall, power rates increased with longer tests, larger sample sizes and DIF magnitudes, and when groups had equal sample sizes and base rates. It is important to note that the power rates of LRT method were slightly higher than that of Wald method, however, caution is necessary when interpreting the power rates results due to inflated and deflated Type I error rates of methods.

Regarding the DIF type, our study showed that the methods exhibited higher power rates for the intercept only DIF and all parameters DIF conditions compared to the main effect and interactions DIF conditions. These findings align with previous literature indicating that the Wald method performs better when uniform DIF exists (Hou et al., 2014; Svetina et al., 2018; Liu et al.,2019; Ma et al., 2021). As the intercept only DIF and all parameters DIF (by our study design item parameters designed to be against to focal group for both masters and nonmasters) affect both masters and nonmasters at the same direction, we can categorize these DIF types as uniform DIF. Also, for main effect and interactions DIF, as it affects only masters, we can categorize this as nonuniform DIF. It is also important to note that all parameters DIF can also be categorized as nonuniform when the parameters have different signs depending on the cancelation effect. Thus, it is important to consider the DIF type when interpreting the results.

Overall, our study demonstrated the effectiveness of the Wald method in detecting DIF items within the LCDM model. This study contributes to the literature by extending the Wald DIF detection method to the general DCM model by addressing additional critical factors that were previously overlooked. Our study enhances the generalizability of the findings by considering extended simulation conditions with the LCDM model. However, limitations include fixing the number of attributes and quality at a moderate level, which may affect the generalizability of the results. Previous research has shown that the Wald method in detecting DIF tends to produce inflated Type I errors when dealing with poor quality items (Hou et al., 2014; Li & Wang, 2015; Svetina et al., 2018; Liu et al., 2019; Ma et al., 2021). Therefore, future research should integrate item quality into the design to enhance the generalizability of our findings.

References

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, *33*(1), 2-14.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

De La Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595.

Feng, Y. (2021). *Effect Size Measures for Differential Item Functioning in Cognitive Diagnostic Models*. Indiana University.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74(2),* 191–210.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning*. Doctoral dissertation, University of Georgia, Athens, GA.

Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28-54.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 1137.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1-26.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting Differential Item Functioning Using Multiple-Group Cognitive Diagnosis Models. *Applied Psychological Measurement*, *45*(1), 37-53.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963-990.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

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

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267-281.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *43*(1), 88-115.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6:4, 219-262.

Rupp, A., Templin, J. and Henson, R. A. (2010) Diagnostic Measurement: Theory, Methods, and Applications. Guilford Press.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in psychology*, *9*, 696.

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

Zhang, W. (2006). *Detecting differential item functioning using the DINA model*. Doctoral dissertation, The University of North Carolina at Greensboro.

CHAPTER 3

INVESTIGATING CRITERIA FOR ITEM FLAGGING IN DIAGNOSTIC CLASSIFICATION MODELS [2]

[2] Zor, S. and M.J. Madison. To be submitted to *Journal of Educational Measurement*.

Abstract

It is necessary to establish criteria for identifying the degrees to which differential item functioning (DIF) is practically significant in the diagnostic classification models (DCMs) framework. This study focuses on investigating criteria based on the degree to which DIF items impact the classification accuracy in a general DCM framework. A simulation study was conducted to investigate the effects of DIF on classifications and compared with the unsigned area (UA) effect size measure. Results showed that classifications remained robust when DIF involved in only intercept parameter; however, noticeable decreases in accuracy were observed when DIF involved in main effect and interactions or all parameters. The findings indicated that the UA effect size measure was effective in capturing DIF levels when a large proportion of DIF items was present. The findings emphasized the limitations of relying solely on the UA effect size measure for item flagging. The results highlighted that the UA effect size measure, while providing valuable insights into DIF level, may not always align with the actual impact on classifications. Depending on the DIF type, the UA effect size measure flagged items as large DIF, despite the robustness of the classifications. Conversely, the UA effect size measure identified negligible and moderate DIF while more pronounced impact on classifications occurred.

Introduction

In the Diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010), framework, the DIF detection methods used a statistical test to examine whether DIF exists or not (Hou et al., 2014, Li & Wang, 2015; Zhang, 2006). The significance test is an essential step in examining test items for DIF, however, the significance test without effect size measures has been criticized (Camilli, 2006; Cohen, 1994; Kirk, 2007). The significance of the statistical test is affected by sample size (Cohen, 1988); therefore, it is vital to determine if the detected DIF is meaningfully large.

Previous DIF studies have shown that statistical power is dependent on sample size and power increases as the sample size increases (Mazor, Clauser, & Hambleton, 1992; Narayanon & Swaminathan, 1996; Swaminathan & Rogers, 1990). This implies a negligible level of DIF can be detected as significant under the large sample size conditions such as in large-scale assessments. Identifying non-DIF items as DIF might result in removing those unbiased items which in turn might affect the item bank and increase the cost of the item/test development process. Thus, in DIF studies, besides detecting significant DIF, it is also important to have a descriptive statistic that quantifies the magnitude of DIF. As Potenza and Dorans (1995) noted, "to be used effectively, a DIF detection technique needs an interpretable measure of the amount of DIF" (p. 33). The effect size measures can help facilitate practical interpretations of DIF test results and to make decisions about the treatment of DIF items. That's why in many DIF studies, exploring effect size measures is of interest (Jodoin & Gierl, 2001; Kim, Cohen, Alagoz, & Kim, 2007; Suh, 2016).

Several effect size measures have been developed for observed score-based DIF detection methods (Dorans & Kulick, 1986; Holland & Thayer, 1988; Schmitt & Dorans, 1990; Shealy &

Stout, 1993; Zumbo & Thomas, 1997) and IRT model-based DIF detection methods (Raju, 1988; Wainer, 1993). In order to assess the magnitude of DIF, established thresholds are used to flag DIF items as negligible or significant. However, currently available thresholds may not be directly used for DCM-based DIF analysis because of the multidimensional nature of DCMs. In the unidimensional framework, the matching variable for DIF analysis is the observed test score or single ability estimate. However, DCMs assume multidimensional latent classes that classify students into mastery or nonmastery levels of attributes. Therefore, the total score or a single ability estimate may fail to represent the multidimensionality of the latent construct. Instead of using the total score as a matching variable in the DCM framework, matching on attribute profiles results in better Type I error and power rates (Zhang, 2006). Thus, flagging criteria established for unidimensional context may differ for DCMs.

In the DCM framework, there have only been two studies that investigated the effect size measures for DIF detection (Feng, 2021; George & Robitzsch, 2014). These effect size measures are based on the probability differences between groups and showed effective results in differentiating the levels of DIF. However, DIF thresholds in DCMs should also consider the impact of DIF on classification accuracy since classification results determine subsequent inferences and decision making. Paulsen et al. (2020) investigated the impact of DIF on DCM classifications under the DINA model. They reported inadequate classifications under the moderate DIF magnitude conditions (0.10) when group distributions are unequal, and also under the large DIF magnitude conditions (0.30) when group distributions are equal. They explored the robustness of DCM classifications to several DIF conditions where the presence of DIF was ignored. However, further research is required on DIF influence on DCMs classifications in order to define the magnitude of DIF in terms on classifications so that to propose levels for item

flagging. Furthermore, the DINA model was used in current studies which is one of the most restrictive models that simplify the DCM estimation. The DINA model is a non-compensatory model that requires the mastery of all attributes to endorse an item. However, this simplified structure may not always hold in practice, and the more general DCMs may be more appropriate in different conditions. Since the general DCMs are more complex and have more parameters to estimate, it is necessary to investigate the impact of DIF on DCM classifications under a more general DCM.

Thus, the main goal of this study is to help establish criteria for identifying the practical significance of DIF in DCMs. We explore the impact of DIF on DCM classifications under a more general DCM. By investigating DIF in a general DCM context, the study aims to provide practitioners with criteria based on the degree to which DIF items impact classification accuracy and reliability. In order to provide an extensive guideline for item flagging, we compare the classification accuracies with the recently developed effect size measures for DCMs, unsigned area effect size measure (UA) (George & Robitzsch, 2014). These criteria will help identify the practical significance of DIF in DCMs and guide the flagging of items in practice.

Method

In the DCM Framework, DIF detection methods have been developed based on the significant test (Hou et al., 2014; Li & Wang, 2014) and there have only been two studies that adopted effect size measures to detect DIF (Feng, 2021; George & Robitzsch, 2014). George and Robitzsch (2014) proposed an effect size measure based on the unsigned area (UA) originally introduced by Raju (1990). This effect size measure is based on the weighted difference of the item response functions between groups:

$$UA_j = \sum_{l=1}^{L} w(\alpha_l) \left| P\left(X_j = 1 | \boldsymbol{\alpha}_l, g_1\right) - P\left(X_j = 1 | \boldsymbol{\alpha}_l, g_2\right) \right|, \quad (1)$$

where $w(\boldsymbol{\alpha}_l) = \frac{1}{2}\left(P(\boldsymbol{\alpha}_l | G = g_1) + P(\boldsymbol{\alpha}_l | G = g_2)\right)$, and $P(\boldsymbol{\alpha}_l | G = g_1)$, $P(\boldsymbol{\alpha}_l | G = g_2)$ are the probability of being in latent class l given group 1 and group 2, respectively. They adopted the cut-off values of .059 to distinguish between negligible and moderate DIF, and .088 for moderate and large DIF, which was suggested by Jodoin and Gierl (2001) in the context of the three parameter model. They conducted the DIF analysis in a large-scale assessment in order to check the item parameter invariance between the groups prior to applying multiple group CDMs. However, the effectiveness of the adopted cut-off values, which is based on the IRT framework, requires further investigation to be used for the effect size measure based on multidimensional attribute profiles.

The studies have shown the suggested effect size measures with existing threshold values based on the probability differences between groups showed effective results in differentiating the DIF categories. However, the effect of DIF items on DCMs' primary purpose of examinee classifications require further investigation in order to provide an understanding of how different DIF related scenarios affect classification accuracy and to provide threshold values for item flagging. Additionally, the DINA model was considered in these studies which is a noncompensatory model that requires the mastery of all attributes to endorse an item. Since the general DCMs are more complex and have more parameters to estimate, it is necessary to assess the performance of the effect size measures under a more general DCMs and investigate criteria for identifying the degrees to which DIF is practically significant.

The current study focuses on investigating thresholds based on the degree to which DIF items impact classification accuracy. Adopted from the IRT framework, currently developed

measures in DCM framework use existing threshold values based on the probability differences between groups in order to assess the magnitude of DIF. Although it has been found that various methods exist for effectively detecting DIF and its impact on item parameters, it is unclear how much DIF affects the classification accuracy. It requires further investigation to define a big, medium and small impact of DIF in terms of classifications and how big of DIF causing problems for classifications. We focus on comparing the classification accuracies under a wide range of conditions with general DCMs. Then, we compare the classification accuracy rates with the UA effect size measure in order to define the magnitude of DIF in terms of classifications. For this, we plot the classification accuracies and the effect sizes to get a continuous scale that provides the thresholds of small, medium, and large DIF with respect to classifications.

Simulation Study

Simulation studies were conducted to systematically investigate DIF in a general DCM framework. In the first study, we investigated the performance of the Wald DIF detection method under the LCDM model. Then, we manipulated the simulation conditions for the second study and focused on establishing criteria for identifying DIF based on the degree to which DIF items impact the classification accuracy.

Simulation Study Design

A simulation study was conducted to investigate the effects of several DIF related factors on the classifications. It aims to investigate criteria for item flagging based on the degree to which DIF items impact classification accuracies. The comparison of the classification-based criteria with the UA effect size measure (George & Robitzsch, 2014) was investigated in order to provide an extensive guideline of item flagging in the DCM framework.

We manipulated 4 key factors for a total of 182 conditions: sample size (equal and unequal sample sizes for focal and reference groups), percentage of DIF items (3 levels), DIF size (10 levels) and DIF type (3 levels; intercept only, main effects and interactions only, and all parameters). Crossing these four factors yielded 180 simulation conditions. In addition to these, we included 2 no-DIF conditions that include no DIF items for two sample sizes. For all conditions of the simulation study, we fixed the number of attributes at 4 and the test length at 24 across all simulation conditions. A balanced Q-matrix was used to measure each attribute with eight items (Table 3.1). The attribute tetrachoric correlations were fixed at .50, and the reference group correct response probabilities for masters (ranges between .70 to .80), non-masters (ranges between .20 to .30), partial masters (ranges between .50 to .60) to be the same across conditions. We selected those item response probability ranges to have a test with moderate difficulty. Under all conditions, the generating and estimation model was the LCDM. We conducted 100 replications for each condition. The simulation study was implemented in R (R Core Team). Table 3.2 summarizes the manipulated factors for this study and these simulation conditions are described in detail below.

Sample size

Sample sizes were manipulated to have equal and unequal sample sizes for focal and reference groups. We used two levels of sample sizes for each group, and the reference group sample sizes were larger than focal group sample sizes under the unequal sample size conditions. The equal sample sizes for each group were 1000/1000 while unequal sample sizes for reference and focal group were 1500/500, respectively.

Percentage of DIF items

The percentages of DIF items for each condition had four levels. We simulated 0, 1/8 (3 items), 1/4 (6 items), and 1/2 (12 items) of the items to be DIF items. These percentages fall within the range of previous DIF studies in DCM framework (Hou et al., 2014; Li & Wang, 2015; Liu et al., 2019; Ma, Terzi, & de la Torre, 2021; Paulsen et al., 2020). We fixed these items over conditions and used the same items as DIF items.

DIF Magnitude

The differences in the LCDM item parameters ($\Delta_{\lambda_{i,0}}, \Delta_{\boldsymbol{\lambda_i}}$) between the focal and reference groups is defined as the DIF size. $\Delta_{\lambda_{i,0}}$indicates the difference in the intercept parameters between the focal group and reference group, and $\Delta_{\boldsymbol{\lambda_i}}$ indicates the difference in the main effects and possible interaction terms for item *i* between the focal and reference groups. Most of the previous studies used .05 and .10 as small and large DIF (Hou et al., 2014; Li & Wang, 2015; Svetina et al., 2018; Zhang, 2006), while some of the others also examined additional levels of .15 and .20 (Feng, 2021; Paulsen et al., 2020). In order to investigate the effect of DIF magnitude, ten levels of DIF size were chosen: $\Delta_{\lambda_{i,0}} = \pm .1k$, $\Delta_{\boldsymbol{\lambda_i}} = \pm .1k$, $k \in \{1,2,\dots,9,10\}$. The focal group's item parameters were obtained by subtracting DIF sizes from the reference group's item parameters. Note that DIF magnitudes used in the previous studies reflected the probability differences because the item parameters were guessing and slipping parameters. The DIF magnitudes used in this study are corresponding values in the logit scale.

DIF Type

We manipulated the DIF types based on the LCDM item parameters where we introduced DIF to only intercept, only main effects and interactions, and all parameters. For intercept only conditions, the focal group's intercept parameters were obtained by subtracting DIF sizes from

the reference group's intercept parameters, and the other item parameters didn't change ($\Delta_{\lambda_{i,0}} < 0, \Delta_{\lambda_i} = 0$). For main effect and interactions only conditions, DIF was applied in only for LCDM main effect and interactions ($\Delta_{\lambda_{i,0}} = 0,\ \Delta_{\lambda_i} < 0$). Lastly, we manipulated all item parameters by decreasing the focal groups item parameters relative to reference group item parameters ($\Delta_{\lambda_{i,0}} < 0\ and\ \Delta_{\lambda_i} < 0$). These conditions where the reference group item parameters are larger than focal group's item parameters were selected to reflect the situations where DIF items are against the focal group. Table 3.3 summarizes the combination of DIF sizes and DIF types for this study.

Evaluation Criteria

We evaluated the robustness of the classifications to several DIF conditions with the classification accuracies and the reliabilities. Then, in order to investigate criteria based on the classifications, we compared the classifications with the UA effect size measure adopted for DCMs (George & Robitzsch, 2014).

Classification Accuracy (CA). Classification accuracy rates were investigated to evaluate to what degree the estimated mastery of attributes matches with the true mastery. The classification accuracies were computed for each attribute and the attribute profile. The classification accuracy for an attribute was calculated as the proportion of examinees whose estimated mastery levels for each attribute were the same as their true mastery level (Attribute Classification Accuracy; ACA). The classification accuracy for the attribute profile was calculated as the proportion of the examinees whose estimated mastery profiles are the same as their true mastery profiles (Profile Classification Accuracy; PCA). We calculated the classification accuracy rates for both focal group and reference group. The no-DIF conditions were used as a baseline to show the impact of DIF on classifications.

Reliability. Reliability refers to the degree to which an examinee's mastery classification is consistent over a series of repeated hypothetical observations (Templin & Bradshaw, 2013). In this study, we used Templin and Bradshaw (2013)'s tetrachoric correlation-based metric to calculate classification reliability. Higher values indicate a more consistent classification.

Then, in order to investigate the DIF flagging criteria based on the classifications, we compared the classification accuracies with the unsigned area (UA) effect size measure (George & Robitzsch, 2014). For this, we plot the classification accuracies and the effect sizes to get a continuous scale. This comparison provides information on how big of DIF causing classification problems and helps to define large, medium, and small impact of DIF in terms of classifications.

Simulation Study Results

We first present the results for classification accuracy and reliability. We report on the results of the classification accuracy rates for focal group and the differences in classification accuracy rates between focal and reference groups. The differences in CAs between groups were reported in order to investigate the DIF impact across the groups. We then report on results from the comparison of the CAs and the UA effect size measure.

Classification Accuracy

Classification Accuracy: Focal Group

Table 3.4 and 3.5 show the classification accuracy rates of focal group across the DIF percentages, DIF magnitude and DIF types under the equal and unequal sample size conditions, respectively. Figures 3.1 and 3.2 present the visualization of these results. Under the null conditions where no-DIF was introduced, the profile classification accuracy (PCA) was 0.649 and 0.650 (equal and unequal sample sizes, respectively) and the attribute classification accuracy

(ACA) was 0.899. The DIF type, DIF magnitude and DIF percentages are three main factors that had an impact on the classifications.

DIF type, including DIF introduced to only intercept parameters, main effect and interaction parameters, or all item parameters, had a noticeable impact on classification accuracies for focal group. The classifications remained relatively robust for DIF introduced to only intercept parameter but were impacted when DIF was introduced to main effect and interaction parameters or all item parameters. Overall, the PCA ranged from 0.652 to 0.573 for DIF conditions involving only the intercept parameter, from 0.647 to 0.511 for DIF conditions involving main effect and interaction parameters, and 0.649 to 0.490 for DIF conditions involving all item parameters (intercept, main effect, and interaction parameters). On the other hand, the ACA ranged from 0.899 to 0.871 for DIF conditions involving only the intercept parameter, from 0.898 to 0.847 for DIF conditions involving main effect and interaction parameters, and 0.898 to 0.840 under the DIF conditions involving all item parameters. When DIF was present in only intercept parameters, the PCA and ACA rates decreased by up to 5% and 2% respectively under equal sample sizes, and 8% and 3% respectively under unequal sample sizes when DIF magnitude was at its highest. Under DIF conditions involving either main effect and interactions or all parameters, the PCA rates decreased by up to 5% for small and medium DIF percentages (where 1/8 and 1/4 of the items were DIF items), as well as for large DIF percentages (where 1/2 of the items were DIF items) when DIF magnitudes smaller than 0.4. The drop in PCA rates were between 5% to 10% when DIF conditions involving main effect and interactions or all parameters, along with large DIF percentages and DIF magnitudes between 0.4 and 0.7. Notably, the PCA rates exhibited a significant decrease (greater than 10%) when half of the items were DIF items and the DIF magnitude exceeded 0.7. For example, the PCA rates

decreased by 14% and 16% when DIF was introduced to all parameters, half of the items exhibited DIF, and the magnitude of DIF was 1 under the equal and unequal sample size conditions, respectively. On the other hand, for conditions involving main effect and interactions or all parameters, the ACA rates demonstrated a decrease of up to 1% when 1/8 of the items exhibited DIF. The ACA rates for equal and unequal sample sizes further declined to 2% and 3% respectively when 1/4 of the items were DIF items, and to 5% and 6% respectively when half of the items were DIF items.

The magnitude of DIF also influenced the classification accuracies. Higher magnitudes of DIF had a more significant impact on classification accuracies. The classifications exhibited a more pronounced decrease compared to the conditions without DIF when DIF was introduced to main effect and interaction parameters or all parameters. For example, when 50% of the items were DIF items under the equal sample sizes and DIF was introduced to main effect and interaction parameters, the focal group's PCA decreased from 0.649 for conditions without DIF to approximately 0.601, 0.551 and 0.529 as the DIF magnitude increased to 0.4, 0.8, and 1, respectively. The decrease in PCA rates were about 5%, 10% and 12% for corresponding conditions, respectively. The corresponding decrease in ACA rates was relatively small, with approximate reductions of 2% (to 0.882), 4% (to 0.863) and 4% (to 0.855) for the respective conditions.

The proportion of DIF items in the test also affected the robustness of classifications. As the percentage of DIF items increased, there was a corresponding decrease in the accuracy of the classifications, particularly with an increase in the magnitude of DIF. When DIF involved only intercept parameter, the classification rates remained relatively consistent across the various conditions, regardless of the DIF percentages. For example, when DIF magnitude was at highest

(1), the PCA was 0.639, 0.628, and 0.597 for small (1/8), medium (1/4), and large (1/2) DIF percentage conditions, respectively. The decrease was about 1%, 2% and 5% for the corresponding DIF percentages compared to no-DIF conditions. On the other hand, when DIF involved main effect and interaction parameters or all parameters, the decrease on the classification rates became more evident with an increase in the percentage of DIF items. For instance, when 0.8 DIF magnitude was introduced to all parameters under the equal sample size conditions, the PCA compared to no DIF conditions decreased by 2% (to 0.627), 5% (to 0.597), and 11% (to 0.538) for small, medium, and large DIF percentage conditions, respectively.

The equal and unequal sample sizes for reference and focal groups didn't significantly affect the classification accuracies. We found that equal sample sizes resulted in slightly more robust classifications compared to unequal sample sizes. Specifically, the PCA differences between equal and unequal sample sizes were 0.01 when DIF magnitude exceeded 0.8 and 1/4 of the items exhibited DIF, and it was 0.02 when half of the items exhibited DIF.

Classification Accuracy: Focal vs Reference Group

Figure 3.3 shows the PCA rates of the reference group across the DIF percentages, DIF magnitude and DIF types under the equal and unequal sample size conditions and Figure 3.4 shows the differences in profile level classification accuracies (PCA) between reference and focal groups. The item parameters of the focal group, including the DIF items, were generated by subtracting the DIF magnitude from the reference group's item parameters. The effect of DIF items on the reference group's classification accuracies were also examined across the conditions (see Figure 3.3). Overall, the results revealed that the presence of DIF items across the conditions did not have a significant impact on the classification accuracies of the reference group. The largest decrease in PCA rates observed was 0.02 when compared to the no-DIF conditions. As a

result, the reference group exhibited higher classification accuracies compared to the focal group. This led to noticeable differences in the classification accuracies between the two groups, as illustrated in Figure 3.4. Thus, the factors that impacted the classification accuracies of the focal group also influenced the differences in classification accuracies between groups in much the same way.

In the DIF type conditions, the classification rates didn't change significantly across the reference and focal groups when the DIF was introduced to only intercept parameter. The largest PCA difference between groups was 0.03 when the DIF magnitude was 1 and half of the items exhibited DIF under the equal sample size conditions, and that of 0.07 under the unequal sample size conditions. Under the conditions where DIF was introduced to main effect and interactions or all item parameters, the differences in PCA between groups were approximately 0.03, 0.06 and 0.13 for equal sample size conditions, and 0.03, 0.07, and 0.16 for unequal sample size conditions when the DIF percentage was 1/8, 1/4, and 1/2, respectively.

In response to DIF magnitude, there was no significant change in classification accuracies across groups for cases where DIF type was intercept only or when the DIF percentage was small or medium. On the other hand, when DIF was introduced to main effect and interactions or all parameters, the differences increased as the DIF magnitude increased under the large DIF percentage (50%) conditions. The PCA rates exhibited minimal variations across the groups when the DIF magnitude ranged from 0.1 to 0.4. However, the differences in PCA rates ranged from 0.05 to 0.10 when the DIF magnitude ranged from 0.4 to 0.08 and exceeded 0.10 for DIF magnitudes greater than 0.8.

When considering sample sizes, there were no notable differences in classification accuracies across groups under the DIF conditions examined. Specifically, when the sample sizes

of the reference and focal groups were equal, the classification accuracies of the reference group showed a slight decrease compared to the no-DIF conditions. However, when the sample sizes were unequal (with a smaller sample size for the focal group), the reference group's classification accuracies remained relatively unchanged.

Reliability

Table 3.6 and 3.7 present the classification reliability results across various DIF percentages, DIF magnitude and DIF types under the equal and unequal sample size conditions, respectively. These results are further depicted in Figure 3.5, providing a visual representation of the results.

In the absence of DIF, the classification reliability was 0.803 for equal sample size conditions and 0.804 for unequal sample size conditions. Overall, across the various DIF conditions, the classification reliabilities followed a similar trend to the classification accuracy. Specifically, as the DIF magnitude and DIF proportions increased, the reliabilities demonstrated a gradual decrease. This effect was more pronounced when DIF was introduced to main effect and interactions or all parameters. Nevertheless, it is important to note that the impact of DIF on classification reliabilities was relatively small compared to classification accuracy, with a maximum decrease of 0.07 observed.

From Table 3.6, the reliabilities for the intercept only DIF conditions ranged from 0.783 to 0.804. The reliabilities exhibited a decreasing trend with increasing DIF percentage. Specifically, the average reliabilities were 0.802 for small DIF percentages, 0.800 for medium DIF percentages and 0.796 for large DIF percentages. For the DIF conditions involving main effect and interaction parameters or all parameters, the reliabilities ranged from 0.731 to 0.802. Similarly, these reliabilities also showed a decreasing pattern as the DIF percentage increased.

The average reliabilities for these conditions were 0.795 for small DIF percentages, 0.786 for medium DIF percentages and 0.765 for large DIF percentages.

Table 3.7 shows the reliabilities obtained under unequal sample size conditions, where the sample size for focal group is 500 and for reference group is 1500. The results showed similarities to the equal sample size conditions, with either no increase or a slight increase in the reliabilities for each corresponding condition. The lowest reliability observed under unequal sample size conditions was .767 when the DIF magnitude was 1 and half of the items exhibited DIF. In comparison, the corresponding reliability for the equal sample size condition was .731.

Classification Accuracy vs Unsigned Area (UA) Effect Size Measure

As part of our investigation for identifying levels of DIF based on classifications in DCMs, we compared the classification accuracies and the UA effect size measure. In Figure 3.6, we plotted the profile classification accuracy (PCA) for focal group and the UA effect size measure across three different DIF percentage conditions (small, medium, large %). The dotted lines in the figure represent the cut-off values of the UA effect size, distinguishing between negligible and moderate DIF (0.059) and between moderate and large DIF (0.088). The three panels in Figure 3.6 correspond to different DIF types: intercept only, main effect and interactions, and all parameters. Additionally, Table 3.8 and 3.9 present the UA effect size estimates for different DIF percentages, DIF magnitudes, and DIF types under the equal and unequal sample size conditions, respectively.

From Figure 3.6, we observed the levels of DIF items flagged by the UA effect size across the conditions. The UA effect size estimates falling in the negligible, moderate, and large effect size category among the conditions were impacted by the DIF type, DIF magnitude and DIF percentages. As the DIF magnitude and percentage of DIF items increased, the UA effect

size estimates also increased. However, the equal and unequal sample sizes didn't have a significant impact on the UA effect size estimates.

In the small (1/8) and medium (1/4) DIF proportion conditions, the UA effect size didn't differentiate all three levels of DIF, only doing so when large percentage (50%) of DIF items were present. Specifically, in the small DIF percentage conditions (1/8), the UA effect size didn't distinguish the levels of DIF, and the corresponding PCA drop compared to no-DIF conditions was 0.01, 0.03, and 0.03 for the intercept only, main effect and interaction parameters, and all parameters DIF type conditions, respectively. Under the medium percentage of DIF conditions (1/4), the UA effect size only distinguished between negligible and moderate DIF when the DIF type was intercept only and all parameters. The intercept only DIF items with a magnitude greater than 0.7 and all parameters DIF items with a magnitude greater than 0.6 were flagged as moderate DIF. The average PCA drop under these conditions was about 0.02 and 0.06 for intercept only and all parameters DIF type conditions, respectively.

The UA effect size differentiated the levels of DIF when a large percentage of DIF items (50%) was present. In the intercept only DIF type conditions, items were flagged as negligible DIF when the DIF magnitude was less than 0.5, as moderate DIF when the DIF magnitude was between 0.5 and 0.8, and as large DIF when the DIF magnitude was greater than 0.8. The corresponding classification accuracy drop for those conditions was not significant, ranging between 0.01 to 0.05. This indicates that even though the UA effect size indicated large DIF when a DIF magnitude of 1 was introduced to intercept parameters, the classifications remained robust, with a maximum decrease of 0.05.

In the main effect and interaction DIF type conditions with large DIF percentages, DIF items with a magnitude smaller than 0.7 were flagged as negligible, and a magnitude between 0.7

and 1 was flagged as moderate DIF. The UA effect size estimates were smaller when DIF was introduced to main effect and interactions compared to intercept only DIF conditions (see Table 3.8 and 3.9). This can be attributed to the fact that in intercept only DIF conditions, DIF affects both nonmasters and masters, whereas in main effect and interactions DIF conditions, DIF influences only masters. From the computation of the UA effect size, it is anticipated to observe smaller UA estimates when DIF is introduced to main effect and interactions. On the other hand, the main effect and interaction DIF type conditions exhibited greater decreases in classification rates compared to intercept only DIF conditions. Specifically, the corresponding classification drops for main effect and interaction DIF type conditions were about 0.05 when DIF was flagged as negligible and 0.11 and 0.13 when DIF was flagged as moderate under equal and unequal sample sizes, respectively.

Lastly, under the large DIF percentage conditions where DIF was introduced to all parameters, DIF items with magnitudes smaller than 0.3 were flagged as negligible DIF, magnitudes between 0.3 and 0.6 were flagged as moderate DIF, and magnitudes exceeding 0.6 were flagged as large DIF. The average classification drops when DIF items were flagged as negligible is 0.03, as moderate is 0.07, and as large is 0.12 for equal sample size conditions.

Simulation Study Conclusions

The simulation study initially focused on analyzing the impact of DIF on DCM classifications and reliability. We examined the robustness of the classifications to several DIF conditions by assessing classification accuracies and reliability. To further investigate the criteria for flagging DIF items based on classifications, we compared classification accuracies with the unsigned area (UA) effect size measure adopted for DCMs (George & Robitzsch, 2014). By plotting classification accuracies against effect sizes, we obtained a continuous scale that helped

determine the extent of DIF impact on classifications, enabling to distinguish between small, medium, and large effects.

Firstly, the study investigated the classification accuracies of the focal group under different DIF conditions and sample sizes. The results demonstrated that the type, magnitude, and percentage of DIF were influential factors in classification accuracy. The classifications remained relatively robust when DIF was introduced to only intercept parameter. However, when DIF was introduced to main effect and interactions or all item parameters, noticeable decreases in classification accuracy rates were observed, particularly with higher magnitudes of DIF. The proportion of DIF items in the test also impacted classifications robustness, with higher DIF percentages leading to decreased accuracy, especially when combined with larger DIF magnitudes. Interestingly, equal and unequal sample sizes for reference and focal groups did not significantly affect classification accuracies, although slightly more robust classifications were observed under equal sample sizes. Overall, significant decreases in classification accuracy were observed for the focal group when DIF was introduced to main effect and interactions or all parameters, particularly with high DIF magnitudes and large DIF percentages. These findings align with existing literature on DCMs that highlights the impact of DIF on classification accuracies (Paulsen et al., 2020).

The study examined classifications at both the profile level (PCA) and individual attribute level (ACA). The PCA and ACA exhibited similar trends across DIF conditions (see Figures 3.1 and 3.2). However, while the classifications at the individual attribute level were robust to DIF conditions, the profile level classifications were not. ACA rates consistently achieved above 80% classification accuracy, indicating higher accuracy in capturing individual attribute information. The decrease in the attribute level classification rates under DIF conditions

compared to no-DIF conditions was minimal, reaching a maximum decrease of 0.06. In contrast, the PCA rates showed larger drops ranging from 0.01 to 0.07 under moderate DIF conditions and 0.01 to 0.16 under high DIF conditions. Overall, ACA proved more robust to DIF conditions, with minimal decreases in accuracy compared to corresponding PCA rates. It can be concluded that significant impacts on ACA rates require substantial DIF magnitudes and a considerable number of DIF items, whereas PCA rates decrease notably even in the presence of moderate levels of DIF.

The simulation study also investigated the differences in classification accuracies between focal and reference groups. The presence of DIF items did not considerably impact the classification accuracies of the reference group, resulting in higher classification accuracies for reference group compared to the focal group. Thus, factors that influenced the classification accuracies of the focal group also affected the differences in classification accuracies between the two groups.

Next, the study examined classification reliability under different DIF conditions. The results revealed that as DIF magnitudes and proportions increased, classification reliabilities gradually decreased. However, the impact of DIF on classification reliability was relatively small compared to classification accuracy.

Finally, as part of our investigation into criteria for identifying levels of DIF based on DCM classifications, we compared classification accuracies and the unsigned area (UA) effect size measure. The results showed that UA effect size estimates were influenced by the DIF type, DIF magnitude and DIF percentages, increasing as DIF magnitude and percentage of DIF items increased. The impact of equal and unequal sample sizes on UA effect size estimates was not significant.

To differentiate between negligible and moderate DIF (0.059) and between moderate and large DIF (0.088), we employed cut-off values for the UA effect size measure based on the work of George & Robitzsch (2014). Under the conditions involving small and medium DIF percentages, the UA effect size measure failed to effectively differentiate between the three levels of DIF. DIF items under these conditions were flagged as negligible or moderate by the UA effect size measure, with negligible average drops observed in PCA, ranging from 0.01 to 0.06 across corresponding conditions. Notably, the UA effect size measure successfully differentiated the levels of DIF when a large percentage of DIF items was present. Our results further revealed the influence of DIF type on UA effect size estimates, where intercept only DIF conditions exhibited higher UA effect size estimates compared to conditions where DIF was introduced to main effect and interactions. This discrepancy can be attributed to the fact that intercept only DIF affects both nonmasters and masters, whereas main effect and interactions DIF affects only masters. From the computation of the UA effect size, smaller UA estimates were observed when DIF was introduced to main effect and interactions. These results suggest that the UA effect size better differentiated the levels of DIF when uniform DIF exist (e.g., intercept only DIF or all parameters DIF, where all parameters consistently higher or lower) than nonuniform DIF (e.g., DIF in main effect and interactions or all parameters, where some parameters higher some lower). Moreover, UA effect size estimates were larger when DIF was present in all item parameters compared to intercept only conditions, indicating greater effect size estimates when DIF was present in more item parameters.

The findings revealed that the UA effect size measure effectively captured levels of DIF only under large DIF conditions (i.e., 50% of DIF items present). However, it is important to note that, although the UA effect size measure flagged DIF items as large DIF under certain

conditions (e.g., intercept only DIF type), the classifications remained robust with no significant decreases observed (maximum decrease in PCA: 5%). Conditions where PCA drop exceeded 10% and UA effect size measure flagged DIF items as moderate were observed when DIF was introduced to main effect and interactions or all parameters. These findings highlight two key points: (1) the significance of considering classification robustness to various DIF conditions when deciding on flagging DIF items and (2) the need to reconsider the threshold values of the UA effect size measure, originally adopted from the IRT framework.

Discussion

In the DCMs framework, the inclusion of DIF analysis and effect size measures holds significant importance. DIF analysis enables the identification of potential biases and ensures fairness in diagnostic assessments by detecting items that function differently across diverse groups. This process enhances the accuracy of diagnostic classifications by identifying items that may unfairly advantage or disadvantage certain groups. Furthermore, the incorporation of effect size measures in DIF analysis provides valuable insights into the magnitude and practical significance of DIF. The effect size measures facilitate the interpretation of DIF results and aid in decision-making regarding the treatment of DIF items.

In this study, we focused on investigating the criteria for identifying the practical significance of DIF items in DCMs. Previous research has focused on effect size measures for DIF detection based on probability differences between groups (Feng, 2021; George & Robitzsch, 2014), but the impact of DIF on classification accuracy in DCMs also needs to be considered, as classifications derived from DCMs play a significant role in subsequent inferences and decision-making process. We extend the investigation by exploring the robustness of DCM classifications under various DIF conditions, using a more general DCM framework. Then, we

compared the classification accuracies and previously developed UA effect size measure to investigate the criteria for flagging DIF items.

Results from the simulation study demonstrate the impact of DIF type, magnitude and proportion on classification accuracy and provide insight into the robustness of the DCM classifications under different DIF conditions. More robust classification accuracies at attribute level compared to profile level classification accuracies were found. The findings highlighted the challenges to DCM classifications associated with DIF introduced to main effect and interaction parameters or all item parameters, particularly when DIF magnitudes are high and a large proportion of items exhibit DIF. This emphasizes the need for careful consideration and mitigation strategies for these types of DIF in DCM applications.

Results from the simulation study show that the UA effect size measure successfully captured levels of DIF under certain conditions, particularly when a large percentage of DIF items was present. It demonstrated an expected increase with higher DIF magnitudes and remained reliable across equal and unequal sample size conditions. In George and Robitzsch's study (2014), they didn't explore the effectiveness of the UA effect size measure with the adopted threshold values via simulations. So, in our study, examining the UA effect size measure for identifying DIF levels via simulations adds to the growing body of literature on DIF in DCMs. The findings regarding the influence of DIF type, DIF magnitude, and sample size on the UA effect size estimate align with literature. Although the study highlights the effectiveness of the UA effect size measure under large DIF percentage conditions, suggests caution when applying it to small and medium DIF scenarios.

Results from comparing classification accuracies and the UA effect size measure highlighted some interesting findings that require to be cautious when interpreting the UA effect

size measure. In conditions involving only intercept parameters, despite being flagged as large by the UA effect size estimate, the classifications remained robust. Conversely, in the main effect and interaction DIF conditions, the UA effect size estimates were smaller and primarily identified negligible and moderate DIF, yet the classification accuracies exhibited a greater decrease. These findings highlight the importance of considering specific DIF types when interpreting the UA effect size results and understanding their implications for DCM classifications.

While the classifications remained robust to most DIF conditions and the UA effect size measure offers valuable insights into the presence and impact of DIF in DCMs, it is important to acknowledge certain limitations of this study. The findings related to the classification accuracies and the UA effect size measure's performance were specific to conditions and factors employed in the study. DCM classifications and its effect on DIF quantification may vary based on the rest of the assessment such as the number of items, item discrimination, and the distribution of skill attributes etc. Thus, the generalizability of these findings to different assessment contexts should be approached with caution.

Although the classifications and the UA effect size measure demonstrated consistent performance across equal and unequal sample sizes, sample size limitations should also be considered. In cases of small sample sizes, the reliability of the UA effect size estimate may be compromised, leading to less precise and potentially unstable results. Further research is necessary to evaluate the robustness and generalizability of the UA effect size measure across diverse sample size settings.

Additionally, the inability of the UA effect size measure to differentiate the levels of DIF under small and medium DIF percentages is a notable limitation. This suggests reduced

sensitivity when the proportion of DIF items is low. One possible explanation for this limitation is that the UA effect size relies on between-group differences in success probabilities. In the conditions where the proportion of DIF items is relatively low, the between-group differences may not be pronounced enough to reach a threshold that distinguishes different levels of DIF. Reconsidering the cut-off values for differentiating DIF levels may be necessary to overcome this limitation.

It is important to be mindful of these limitations when interpreting and applying the UA effect size measure in practice. Researchers should also consider the impact of DIF on DCM classifications to enhance the accuracy and comprehensiveness of DIF detection in DCMs.

DIF magnitude and impact analysis play a crucial role in determining the practical significance of DIF. However, studies investigating DIF, particularly those focusing on effect size measures, remain limited in DCM literature. Therefore, further research is needed to establish optimal cut-off values for DIF in DCMs. By addressing these important issues, our study contributes to the development of practical guidelines for identification and interpretation of DIF in DCMs.

References

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement*, *23*(4), 355-368.

Feng, Y. (2021). *Effect Size Measures for Differential Item Functioning in Cognitive Diagnostic Models*. Indiana University.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure In Wainer H & Braun HI (Eds.), Test validity (pp. 129–145). Hillsdale, NJ, US.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, *14*(4), 329-349.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of educational measurement*, *44*(2), 93-116.

Kirk, R. E. (2007). Effect magnitude: A different focus. *Journal of statistical planning and inference*, *137*(5), 1634-1646.

Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28-54.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, *10*, 1137.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting Differential Item Functioning Using Multiple-Group Cognitive Diagnosis Models. *Applied Psychological Measurement*, *45*(1), 37-53.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443-451.

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

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267-281.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied psychological measurement*, *19*(1), 23-37.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Rupp, A., Templin, J. and Henson, R. A. (2010) Diagnostic Measurement: Theory, Methods, and Applications. Guilford Press.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67-81.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement*, *53*(4), 403-430.

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in psychology*, *9*, 696.

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

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 123–135). Hillsdale, NJ: Erlbaum.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model*. Doctoral dissertation, The University of North Carolina at Greensboro.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

CHAPTER 4

DISCUSSION

Diagnostic classification models have shown great promise in educational assessments to provide fine-grained diagnostic feedback about examinees' knowledge across different attributes (Bradshaw et al., 2014; Madison & Bradshaw, 2018; Rupp, Templin, & Henson, 2010). By modeling the relationship between item responses and latent attributes, DCMs enable the identification of specific areas of strength and weaknesses for each student. However, further research is necessary to explore and refine the application of DCMs, including the investigation of differential item functioning (DIF). Although several DIF detection methods proposed (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Ma et al., 2021), one of the limitations of the DIF DCM literature is that the findings haven't been generalized to the LCDM. Additionally, more research should focus on practical implications of DCM-based DIF analysis. This includes examining the effects of DIF on decision-making such as classification accuracy and the practical significance of DIF. This dissertation focused on (1) investigating the performance of the Wald DIF detection method (Hou et. al., 2014) when extended to the general LCDM framework and (2) investigating the criteria for identifying the practical significance of DIF items based on the DCM classifications and unsigned area (UA) effect size measure (George and Robitzsch, 2014).

Differential Item Functioning in DCMS

DIF assessment plays a crucial role in evaluating the fairness and validity of educational assessments. While the traditional approach to DIF assessment is primarily based on the IRT framework, DCMs has introduced some key differences in the assessment process. One

fundamental distinction between the DCM and IRT framework lies in the underlying attributes being measured. In IRT, the latent traits are typically unidimensional and continuous, assuming a single underlying construct. On the other hand, DCMs allow for the measurement of multiple attributes, providing a more detailed representation of a test-taker's knowledge or skills. These attributes are often binary, indicating whether a test-taker possesses a particular attribute or not. Due to these differences in the frameworks, traditional DIF assessment methods designed for the IRT framework may not be directly applicable to data that follows DCMs. The IRT based DIF methods typically rely on comparing item responses between groups after matching them on a single trait to identify potential DIF items. These methods assume a unidimensional latent trait and may not effectively capture multidimensional attributes assessed by DCMs.

In the DCM framework, several DIF assessment methods have been developed to account for the multidimensionality and binary nature of the underlying attributes (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Ma et al., 2021; Zhang, 2006). These methods include observed score methods such as the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) using attribute profiles as matching variable. Another approach, proposed by Hou et al. (2014), is the Wald test for detecting DIF in DCMs. However, a limitation of previous approach is that the estimates of attribute profiles, used as the matching criterion, may be influenced by DIF items within the test. compromising the validity of subsequent DIF assessment.

To address this issue, Hou et. al. (2014) employed the Wald method to identify DIF in the DINA model. They conducted separate calibrations of the DINA model for the reference and focal groups, allowing for more accurate assessment of DIF by eliminating contamination in item parameters and attribute estimates caused by DIF. This approach proved to be efficient in several

studies (Hou et al., 2014; Liu et al., 2019; Ma et al., 2021; Svetina et al., 2018). However, most DIF studies in DCM literature have predominantly focused on the DINA model, which is one of the most restrictive DCMs. The DINA model assumes a specific theory about attributes, which may not accurately represent the complexity of real-world educational assessment data. Ma et. al. (2021) developed a multiple-group generalized deterministic inputs, noisy "and" gate (MG-GDINA) model and investigate the performance of Wald test in detecting DIF under MG-GDINA model. By extending the Wald method to the LCDM framework, Study 1 aimed to address this limitation and evaluate the performance of the method under a more general DCM framework.

The proposed DIF detection methods in the DCM framework use statistical tests to determine the presence of DIF (Hou et al., 2014; Li & Wang, 2015; Zhang, 2006). However, relying solely on significance tests may lead to the identification of items with insignificant DIF, limiting the interpretability of the results (G´omez-Benito et al., 2013, Zhang, 2006). To address this limitation, it is crucial to incorporate effect size measures when examining DIF. Effect size measures provide valuable information about the magnitude and practical significance of DIF, enhancing the interpretability of the results.

In IRT framework, established thresholds are used to flag DIF items as negligible or significant (Zwick & Ercikan, 1989). However, the multidimensional nature of DCMs requires careful consideration when applying thresholds established for unidimensional framework. Additionally, DIF thresholds in DCMs should also consider the impact of DIF on classification accuracy, as subsequent inferences and decision making rely on classification results. The effect of DIF on the accuracy of attribute classification is a crucial aspect to be considered while determining DIF thresholds in DCMs. Balancing the detection of significant DIF with its

practical implications for classification accuracy ensures that DIF analysis provides meaningful and actionable information for assessments. In response to this concern, in Study 2, we focused on investigating the criteria for item flagging by considering DCM classifications and unsigned area (UA) effect size measure (George and Robitzsch, 2014).

Study 1: Investigating Wald Method in DIF Detection in the LCDM

In this study, we focused on evaluating the performance of the Wald DIF detection method extended to the LCDM model and compared it with the LRT method. The results of the simulation study indicated that the Wald method effectively controls Type I errors across different conditions, while the LRT method shows inflated Type I errors in conditions with small and unequal sample sizes. The findings highlighted the more conservative nature of the Wald method and its superiority in controlling Type I errors even under small and unequal sample sizes, making it a reliable choice for DIF detection.

Regarding power rates, both the Wald and LRT methods performed well in detecting DIF items when the DIF magnitude is large. However, power rates were relatively low for both methods in conditions with smaller DIF magnitudes. Power rates increased with larger sample sizes and DIF magnitudes, as well as when groups had equal sample sizes and base rates. Additionally, it is important to interpret the power rates of the LRT method with caution due to its inflated Type I error rates.

Our study also examined the impact of different DIF types on the performance of the DIF detection methods. We found that intercept only DIF and all parameters DIF conditions exhibited higher power rates compared to main effect and interactions DIF conditions. These results align with previous literature, highlighting that the Wald method performs better when uniform DIF

exists and emphasize the importance of considering the specific type of DIF when interpreting the findings.

Study 2: Investigating DIF Effect Size Measure in DCMs

In Study 2, we conducted a simulation study to explore the effects of DIF on DCMs classifications and to examine criteria for identifying DIF items based on classification accuracies. The findings highlighted the influence of DIF type, magnitude, and proportion on classification accuracies. Classifications remained robust when DIF affected only the intercept parameter, but noticeable decrease in accuracy were observed when DIF affected main effect and interactions or all item parameters, especially with higher DIF magnitudes and larger proportions of DIF items. Sample size imbalances between the reference and focal groups did not significantly affect classification accuracies.

The results also showed that the classifications at the attribute level were pretty robust to various DIF conditions even in the presence of half of the DIF items with the magnitude of 1 in the test. The profile level classifications were smaller than attribute level classifications and were not robust to DIF in large DIF conditions, where the decrease exceeded 10%. We investigated the effect of different DIF conditions on classifications in order to inform decision making on quantifying the magnitude of DIF. Because detection and quantification of DIF is only really meaningful to the extent that it affects DCM classification decisions.

The DIF conditions designed to be against the focal group and the presence of DIF items had minimal impact on the classification accuracies of the reference group. In terms of classification reliability, the study found that as DIF magnitudes and proportions increased, classification reliability gradually decreased, although the impact on reliability was relatively small compared to classification accuracy.

The results of the study also indicated that the UA effect size measure was effective in capturing DIF levels under specific conditions, particularly when a large proportion of DIF items was present. As expected, the UA effect size measure demonstrated larger values with higher DIF magnitudes and showed reliable estimations across equal and unequal sample size conditions. However, it is important to exercise caution when interpreting the UA effect size measure. In situations where only intercept parameters were affected by DIF, the UA effect size measure flagged them as large. However, despite the large effect size estimates, the classifications remained robust. On the other hand, in DIF conditions involving main effect and interaction parameters, the UA effect size estimates were smaller and primarily identified negligible and moderate DIF. However, despite the smaller effect sizes, it had a more substantial impact on classification accuracies, leading to greater decreases. These findings underscore the significance of considering specific DIF types when interpreting the UA effect size results and their implications for DCM classifications.

These results highlight the nuanced nature of DIF and its impact on classification accuracy. The UA effect size measure, while providing valuable insight into DIF levels, may not always align with the actual impact on classifications. The differences observed in the effect sizes and their corresponding classification accuracies suggest that the UA effect size measure should be interpreted in conjunction with other factor, such as the specific DIF item parameters involved, to gain a comprehensive understanding of the practical significance on DIF in DCMs.

By identifying these patterns, the study underscores the need for careful interpretation and contextualization of effect size measures in DIF analysis. Researchers and practitioners should be cautious when solely relying on effect size estimates to determine the significance of DIF and its implications for DCM classifications. Instead, a comprehensive examination of the

specific DIF types, their magnitudes, and their impact on DCM classifications is necessary for a more accurate assessment of DIF in DCMs.

Educational Significance

The findings of these studies enhance our understanding of DCM based DIF assessment and have significant implications for educational assessment practices. The extension of the Wald method to the LCDM framework and the consideration of additional critical factors in Study 1 contribute to the generalizability and applicability of DCM-based DIF assessment. The Wald method proves to be effective in detecting DIF items within LCDM models and makes it a reliable choice even under small unequal sample sizes.

The findings from Study 2 emphasize the practical implications of DIF in DCMs and contribute to our understanding of the impact of DIF on classification accuracy and reliability. By detecting and addressing DIF, these studies promote fairness in diagnostic assessments. DCM classifications play a crucial role in providing accurate information about examinees' abilities and skills, which, in turn, influences decisions about educational placement and intervention strategies. Understanding the impact of DIF on classifications and reliability is essential for ensuring valid and meaningful inferences from assessment results. Furthermore, the examination of effect size measure, such as the unsigned area (UA) effect size measure, contributes to the DCM literature by providing insights into the magnitude and practical significance of DIF. By interpreting both results from effect size measures and classification accuracies, researcher and practitioners can make more informed decisions regarding the treatment of DIF items and identifying levels of DIF. The findings contribute to the development of practical guidelines for identifying and interpreting DIF in educational assessments, ultimately leading to more accurate and equitable educational outcomes for students.

Future Research

Future research can investigate deeper the robustness of DCM classifications under different DIF conditions. This includes exploring the impact of various factors such as item discrimination, number of items, and the distribution of skill attributes on classification accuracy. Understanding the interaction between these factors and DIF will contribute to the development of more accurate and reliable DIF detection and quantification methods. Additionally, while the UA effect size measure has shown promise in capturing DIF levels under certain conditions, further research is needed to refine and improve its effectiveness. This includes exploring alternative effect size measures that may better differentiate between DIF levels, particularly in scenarios with small and medium DIF percentages. Additionally, investigating the influence of different cut-off values for effect size measures can help establish more robust criteria for identifying practical significance of DIF.

Table 2.1. *Summary of Simulation Conditions for Simulation Study I*

Manipulated Factors		Value	
Base Rate for groups	Equal	(.50, .50)	
	Unequal	(.75, .25) (higher for R)	
Test Length	Short, Long	24, 48	
Sample Sizes	Equal	$N_R = 500$, $N_F = 500$ $N_R = 1000$, $N_F = 1000$	
	Unequal	$N_R = 750$, $N_F = 250$ $N_R = 1500$, $N_F = 500$	
Percentages of DIF items		0, 1/8, 1/4, 1/2	
DIF Magnitude	Small	$\lvert\Delta_{\lambda_{i,0}}\rvert = .3$, $\lvert\Delta_{\boldsymbol{\lambda_i}}\rvert = .3$	
	Medium	$\lvert\Delta_{\lambda_{i,0}}\rvert = .5$, $\lvert\Delta_{\boldsymbol{\lambda_i}}\rvert = .5$	
	Large	$\lvert\Delta_{\lambda_{i,0}}\rvert = .8$, $\lvert\Delta_{\boldsymbol{\lambda_i}}\rvert = .8$	
DIF Type		$\Delta_{\lambda_{i,0}} = (\lambda_{i,0})_F - (\lambda_{i,0})_R$	$\Delta_{\boldsymbol{\lambda_i}} = (\boldsymbol{\lambda_i})_F - (\boldsymbol{\lambda_i})_R$
	Intercept only	–	0
	Main effect and interactions	0	–
	All parameters	–	–

Note. N_R = Reference group sample size; N_F = Focal group sample size; - indicates a negative number, 0 indicates no difference between groups; $\Delta_{\lambda_{i,0}}$ = The difference in the intercept parameters between the focal and reference groups, $\Delta_{\boldsymbol{\lambda_i}}$ = The difference in the main effects and possible interaction terms for item *i* between the focal and reference groups.

Table 2.2. *Q-matrix for the Simulation Study I*

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	1	0	0
6	1	0	1	0
7	0	1	0	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0
11	0	1	1	0
12	0	1	0	1
13	0	0	1	0
14	0	0	1	0
15	0	0	1	0
16	0	0	1	0
17	0	0	1	1
18	1	0	1	0
19	0	0	0	1
20	0	0	0	1
21	0	0	0	1
22	0	0	0	1
23	1	0	0	1
24	0	1	0	1

Table 2.3. *Combination of DIF Sizes and DIF Types for Simulation Study*

DIF Type	DIF Magnitude	$\Delta_{\boldsymbol{\lambda_{i,0}}} = (\boldsymbol{\lambda_{i,0}})_{\boldsymbol{F}} - (\boldsymbol{\lambda_{i,0}})_{\boldsymbol{R}}$	$\Delta_{\boldsymbol{\lambda_i}} = (\boldsymbol{\lambda_i})_{\boldsymbol{F}} - (\boldsymbol{\lambda_i})_{\boldsymbol{R}}$
No DIF	0	0	0
Intercept only	.3	-.3	0
	.5	-.5	0
	.8	-.8	0
Main Effect and interactions	.3	0	-.3
	.5	0	-.5
	.8	0	-.8
All parameters	.3	-.3	-.3
	.5	-.5	-.5
	.8	-.8	-.8

Note. $\Delta_{\lambda_{i,0}}$ = The difference in the intercept parameters between the focal and reference groups, $\Delta_{\boldsymbol{\lambda_i}}$ = The difference in the main effects and possible interaction terms for item *i* between the focal and reference groups.

Table 2.4. *Type I Error Rates for Conditions with No DIF across the Test Lengths, Base Rates and Sample Sizes*

Test Length	Base Rate	Sample Size	$\mathbf{N_R}, \mathbf{N_F}$	Wald	LRT
Short	Equal	Equal	500, 500	0.042	0.06
			1000, 1000	0.047	0.055
		Unequal	750, 250	0.038	0.064
			1500, 500	0.041	0.048
	Unequal	Equal	500, 500	0.045	0.063
			1000, 1000	0.054	0.058
		Unequal	750, 250	0.044	0.08
			1500, 500	0.046	0.058
Long	Equal	Equal	500, 500	0.016	0.054
			1000, 1000	0.03	0.05
		Unequal	750, 250	0.005	0.057
			1500, 500	0.02	0.049
	Unequal	Equal	500, 500	0.013	0.059
			1000, 1000	0.026	0.052
		Unequal	750, 250	0.004	0.055
			1500, 500	0.016	0.053

Table 2.5. *Type I Error Rates across the DIF Conditions with Equal Base Rate and Equal Sample Size Under Short Tests*

			Wald			LRT		
Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All	Int Only	Main+inx	All
500-500	0	0	0.042	0.042	0.042	0.06	0.06	0.06
	1/8	0.3	0.043	0.037	0.041	0.06	0.055	0.053
		0.5	0.038	0.04	0.044	0.054	0.055	0.06
		0.8	0.045	0.046	0.05	0.061	0.065	0.064
	1/4	0.3	0.047	0.042	0.044	0.06	0.057	0.062
		0.5	0.039	0.043	0.046	0.052	0.058	0.062
		0.8	0.043	0.045	0.047	0.059	0.063	0.06
	1/2	0.3	0.036	0.04	0.042	0.052	0.056	0.054
		0.5	0.035	0.045	0.059	0.048	0.062	0.07
		0.8	0.047	0.058	0.064	0.064	0.071	0.073
1000-1000	0	0	0.047	0.047	0.047	0.055	0.055	0.055
	1/8	0.3	0.047	0.035	0.047	0.055	0.043	0.052
		0.5	0.041	0.039	0.039	0.046	0.044	0.049
		0.8	0.056	0.04	0.045	0.062	0.05	0.051
	1/4	0.3	0.051	0.036	0.051	0.057	0.044	0.059
		0.5	0.041	0.042	0.041	0.044	0.047	0.048
		0.8	0.054	0.042	0.047	0.059	0.051	0.053
	1/2	0.3	0.047	0.034	0.044	0.054	0.043	0.051
		0.5	0.048	0.04	0.044	0.055	0.044	0.049
		0.8	0.048	0.053	0.046	0.058	0.057	0.053

Table 2.6. *Type I Error Rates across the DIF Conditions with Equal Base Rate and Unequal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	0	0	0.038	0.038	0.038	0.064	0.064	0.064
	1/8	0.3	0.048	0.041	0.047	0.072	0.066	0.067
		0.5	0.046	0.049	0.05	0.069	0.073	0.073
		0.8	0.038	0.049	0.037	0.063	0.064	0.06
	1/4	0.3	0.044	0.047	0.046	0.078	0.072	0.063
		0.5	0.041	0.043	0.05	0.063	0.074	0.071
		0.8	0.042	0.056	0.051	0.06	0.078	0.072
	1/2	0.3	0.06	0.062	0.054	0.079	0.08	0.069
		0.5	0.04	0.06	0.067	0.068	0.078	0.085
		0.8	0.043	0.066	0.078	0.065	0.073	0.09
1500-500	0	0	0.041	0.041	0.041	0.048	0.048	0.048
	1/8	0.3	0.049	0.037	0.035	0.06	0.05	0.046
		0.5	0.045	0.051	0.045	0.054	0.063	0.056
		0.8	0.046	0.055	0.047	0.059	0.07	0.058
	1/4	0.3	0.048	0.038	0.035	0.057	0.052	0.048
		0.5	0.045	0.049	0.046	0.053	0.057	0.057
		0.8	0.047	0.056	0.063	0.061	0.065	0.069
	1/2	0.3	0.038	0.043	0.038	0.057	0.056	0.045
		0.5	0.048	0.053	0.046	0.052	0.063	0.056
		0.8	0.048	0.053	0.068	0.059	0.058	0.074

Table 2.7. *Type I Error Rates across the DIF Conditions with Unequal Base Rate and Equal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald Int Only	Wald Main+inx	Wald All	LRT Int Only	LRT Main+inx	LRT All
500-500	0	0	0.045	0.045	0.045	0.063	0.063	0.063
	1/8	0.3	0.045	0.052	0.051	0.062	0.075	0.07
		0.5	0.046	0.044	0.047	0.06	0.064	0.069
		0.8	0.044	0.05	0.052	0.066	0.066	0.07
	1/4	0.3	0.041	0.055	0.052	0.062	0.078	0.069
		0.5	0.043	0.046	0.059	0.06	0.062	0.074
		0.8	0.042	0.054	0.05	0.066	0.066	0.066
	1/2	0.3	0.038	0.061	0.052	0.057	0.078	0.069
		0.5	0.044	0.045	0.068	0.062	0.061	0.086
		0.8	0.042	0.058	0.068	0.063	0.065	0.074
1000-1000	0	0	0.054	0.054	0.054	0.058	0.058	0.058
	1/8	0.3	0.047	0.051	0.041	0.055	0.056	0.047
		0.5	0.047	0.045	0.04	0.056	0.056	0.051
		0.8	0.044	0.043	0.048	0.054	0.054	0.06
	1/4	0.3	0.054	0.051	0.043	0.062	0.056	0.051
		0.5	0.048	0.049	0.043	0.056	0.054	0.052
		0.8	0.04	0.046	0.059	0.049	0.055	0.067
	1/2	0.3	0.047	0.055	0.051	0.055	0.063	0.055
		0.5	0.051	0.052	0.038	0.06	0.056	0.046
		0.8	0.042	0.053	0.057	0.053	0.058	0.064

Table 2.8. *Type I Error Rates across the DIF Conditions with Unequal Base Rate and Unequal Sample Size Under Short Tests*

			Wald			LRT		
Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	0	0	0.044	0.044	0.044	0.08	0.08	0.08
	1/8	0.3	0.049	0.055	0.043	0.094	0.096	0.083
		0.5	0.046	0.04	0.044	0.086	0.08	0.088
		0.8	0.045	0.044	0.049	0.082	0.077	0.085
	1/4	0.3	0.047	0.058	0.038	0.084	0.093	0.078
		0.5	0.044	0.046	0.052	0.088	0.093	0.097
		0.8	0.044	0.046	0.048	0.081	0.081	0.086
	1/2	0.3	0.056	0.06	0.039	0.093	0.097	0.073
		0.5	0.05	0.046	0.059	0.083	0.103	0.113
		0.8	0.038	0.065	0.053	0.077	0.106	0.098
1500-500	0	0	0.046	0.046	0.046	0.058	0.058	0.058
	1/8	0.3	0.047	0.048	0.046	0.062	0.061	0.052
		0.5	0.049	0.047	0.048	0.061	0.056	0.065
		0.8	0.051	0.053	0.056	0.064	0.069	0.066
	1/4	0.3	0.05	0.053	0.046	0.059	0.068	0.053
		0.5	0.046	0.045	0.054	0.062	0.054	0.062
		0.8	0.053	0.059	0.058	0.063	0.07	0.067
	1/2	0.3	0.041	0.051	0.044	0.053	0.065	0.057
		0.5	0.053	0.057	0.061	0.067	0.055	0.077
		0.8	0.051	0.081	0.069	0.067	0.08	0.077

Table 2.9. *Type I Error Rates across the DIF Conditions with Equal Base Rate and Equal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
500-500	0	0	0.016	0.016	0.016	0.054	0.054	0.054
	1/8	0.3	0.018	0.015	0.015	0.054	0.055	0.051
		0.5	0.016	0.015	0.017	0.05	0.054	0.054
		0.8	0.016	0.019	0.016	0.054	0.055	0.052
	1/4	0.3	0.018	0.015	0.014	0.053	0.05	0.051
		0.5	0.016	0.016	0.019	0.049	0.054	0.055
		0.8	0.016	0.02	0.017	0.051	0.058	0.053
	1/2	0.3	0.011	0.006	0.008	0.049	0.048	0.053
		0.5	0.009	0.01	0.007	0.051	0.053	0.047
		0.8	0.005	0.009	0.008	0.05	0.058	0.05
1000-1000	0	0	0.03	0.03	0.03	0.05	0.05	0.05
	1/8	0.3	0.029	0.029	0.032	0.044	0.051	0.053
		0.5	0.032	0.035	0.03	0.053	0.055	0.048
		0.8	0.025	0.028	0.028	0.046	0.049	0.047
	1/4	0.3	0.03	0.029	0.032	0.044	0.053	0.053
		0.5	0.029	0.035	0.03	0.055	0.054	0.048
		0.8	0.028	0.027	0.031	0.045	0.05	0.049
	1/2	0.3	0.015	0.012	0.016	0.043	0.052	0.055
		0.5	0.02	0.02	0.016	0.058	0.06	0.051
		0.8	0.011	0.015	0.015	0.04	0.05	0.046

Table 2.10. *Type I Error Rates across the DIF Conditions with Equal Base Rate and Unequal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	0	0	0.005	0.005	0.005	0.057	0.057	0.057
	1/8	0.3	0.005	0.006	0.004	0.055	0.056	0.054
		0.5	0.005	0.004	0.005	0.053	0.061	0.058
		0.8	0.003	0.004	0.005	0.046	0.051	0.054
	1/4	0.3	0.005	0.007	0.005	0.052	0.055	0.054
		0.5	0.005	0.004	0.006	0.051	0.06	0.057
		0.8	0.003	0.005	0.005	0.046	0.053	0.055
	1/2	0.3	0.003	0.005	0.002	0.053	0.057	0.055
		0.5	0.001	0.003	0.003	0.051	0.065	0.055
		0.8	0	0.004	0.004	0.045	0.05	0.062
1500-500	0	0	0.02	0.02	0.02	0.049	0.049	0.049
	1/8	0.3	0.019	0.021	0.02	0.049	0.054	0.051
		0.5	0.022	0.019	0.021	0.053	0.049	0.052
		0.8	0.023	0.027	0.021	0.049	0.058	0.054
	1/4	0.3	0.019	0.022	0.021	0.047	0.053	0.051
		0.5	0.023	0.017	0.021	0.056	0.049	0.05
		0.8	0.021	0.028	0.023	0.049	0.059	0.055
	1/2	0.3	0.01	0.012	0.015	0.048	0.059	0.052
		0.5	0.012	0.008	0.009	0.052	0.047	0.044
		0.8	0.01	0.019	0.018	0.043	0.065	0.062

Table 2.11. *Type I Error Rates across the DIF Conditions with Unequal Base Rate and Equal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald Int Only	Wald Main+inx	Wald All	LRT Int Only	LRT Main+inx	LRT All
500-500	0	0	0.013	0.013	0.013	0.059	0.059	0.059
	1/8	0.3	0.012	0.008	0.011	0.057	0.058	0.05
		0.5	0.01	0.014	0.011	0.05	0.061	0.053
		0.8	0.011	0.011	0.012	0.059	0.053	0.052
	1/4	0.3	0.011	0.009	0.01	0.056	0.059	0.051
		0.5	0.01	0.014	0.009	0.049	0.058	0.054
		0.8	0.012	0.011	0.01	0.056	0.053	0.051
	1/2	0.3	0.005	0.006	0.01	0.056	0.055	0.05
		0.5	0.005	0.009	0.008	0.05	0.058	0.059
		0.8	0.008	0.007	0.007	0.056	0.055	0.053
1000-1000	0	0	0.026	0.026	0.026	0.052	0.052	0.052
	1/8	0.3	0.02	0.025	0.026	0.046	0.052	0.049
		0.5	0.032	0.027	0.025	0.053	0.052	0.05
		0.8	0.024	0.022	0.025	0.047	0.049	0.05
	1/4	0.3	0.021	0.024	0.028	0.049	0.049	0.053
		0.5	0.031	0.026	0.029	0.052	0.052	0.051
		0.8	0.023	0.024	0.025	0.045	0.053	0.048
	1/2	0.3	0.009	0.01	0.014	0.047	0.051	0.053
		0.5	0.018	0.016	0.014	0.06	0.055	0.052
		0.8	0.01	0.013	0.013	0.046	0.057	0.05

Table 2.12. *Type I Error Rates across the DIF Conditions with Unequal Base Rate and Unequal Sample Size Under Long Tests*

			Wald			LRT		
Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	0	0	0.004	0.004	0.004	0.055	0.055	0.055
	1/8	0.3	0.003	0.003	0.003	0.065	0.056	0.059
		0.5	0.004	0.003	0.004	0.06	0.06	0.059
		0.8	0.002	0.003	0.003	0.059	0.055	0.061
	1/4	0.3	0.003	0.003	0.003	0.063	0.056	0.058
		0.5	0.004	0.002	0.004	0.062	0.062	0.062
		0.8	0.002	0.004	0.003	0.058	0.055	0.057
	1/2	0.3	0.004	0.003	0.003	0.058	0.059	0.06
		0.5	0.003	0.003	0.001	0.056	0.06	0.063
		0.8	0.002	0.003	0.001	0.054	0.059	0.07
1500-500	0	0	0.016	0.016	0.016	0.053	0.053	0.053
	1/8	0.3	0.018	0.018	0.02	0.054	0.053	0.053
		0.5	0.019	0.017	0.021	0.056	0.053	0.052
		0.8	0.015	0.017	0.02	0.046	0.051	0.059
	1/4	0.3	0.018	0.019	0.023	0.052	0.055	0.054
		0.5	0.019	0.016	0.022	0.057	0.052	0.054
		0.8	0.015	0.018	0.022	0.048	0.049	0.061
	1/2	0.3	0.009	0.009	0.013	0.051	0.051	0.053
		0.5	0.009	0.006	0.008	0.055	0.05	0.054
		0.8	0.008	0.009	0.011	0.049	0.051	0.057

Table 2.13. *Power Rates across the DIF Conditions with Equal Base Rate and Equal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
500-500	1/8	0.3	0.303	0.157	0.37	0.363	0.223	0.413
		0.5	0.627	0.39	0.587	0.693	0.47	0.63
		0.8	0.987	0.843	0.85	0.99	0.887	0.873
	1/4	0.3	0.282	0.132	0.4	0.342	0.19	0.455
		0.5	0.642	0.367	0.693	0.698	0.425	0.72
		0.8	0.983	0.8	0.882	0.99	0.848	0.908
	1/2	0.3	0.27	0.134	0.401	0.326	0.162	0.452
		0.5	0.65	0.365	0.678	0.702	0.422	0.7
		0.8	0.985	0.793	0.89	0.988	0.838	0.907
1000-1000	1/8	0.3	0.563	0.287	0.493	0.587	0.333	0.507
		0.5	0.977	0.763	0.833	0.98	0.79	0.86
		0.8	1	0.99	0.997	1	0.993	1
	1/4	0.3	0.537	0.307	0.593	0.568	0.327	0.613
		0.5	0.953	0.705	0.882	0.963	0.737	0.898
		0.8	1	0.993	1	1	0.997	1
	1/2	0.3	0.536	0.294	0.608	0.568	0.317	0.628
		0.5	0.957	0.705	0.864	0.964	0.725	0.875
		0.8	1	0.993	0.996	1	0.993	0.997

Table 2.14. *Power Rates across the DIF Conditions with Equal Base Rate and Unequal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	1/8	0.3	0.253	0.08	0.277	0.353	0.12	0.36
		0.5	0.503	0.283	0.447	0.597	0.377	0.533
		0.8	0.917	0.65	0.75	0.94	0.757	0.84
	1/4	0.3	0.208	0.09	0.315	0.317	0.148	0.408
		0.5	0.463	0.23	0.572	0.558	0.345	0.642
		0.8	0.91	0.572	0.813	0.95	0.698	0.867
	1/2	0.3	0.216	0.087	0.284	0.3	0.14	0.382
		0.5	0.474	0.245	0.573	0.578	0.338	0.641
		0.8	0.908	0.565	0.821	0.948	0.681	0.87
1500-500	1/8	0.3	0.423	0.273	0.397	0.48	0.3	0.423
		0.5	0.887	0.58	0.723	0.9	0.65	0.757
		0.8	1	0.96	0.973	1	0.977	0.983
	1/4	0.3	0.407	0.212	0.497	0.463	0.252	0.543
		0.5	0.88	0.538	0.8	0.907	0.605	0.823
		0.8	1	0.94	0.982	1	0.955	0.985
	1/2	0.3	0.41	0.208	0.503	0.469	0.258	0.543
		0.5	0.883	0.523	0.782	0.899	0.582	0.801
		0.8	0.999	0.922	0.967	0.999	0.933	0.972

Table 2.15. *Power Rates across the DIF Conditions with Unequal Base Rate and Equal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald Int Only	Wald Main+inx	Wald All	LRT Int Only	LRT Main+inx	LRT All
500-500	1/8	0.3	0.17	0.087	0.2	0.247	0.137	0.283
		0.5	0.373	0.24	0.48	0.513	0.33	0.527
		0.8	0.883	0.613	0.713	0.93	0.713	0.77
	1/4	0.3	0.153	0.083	0.245	0.23	0.132	0.328
		0.5	0.367	0.222	0.555	0.487	0.318	0.605
		0.8	0.833	0.527	0.78	0.913	0.65	0.812
	1/2	0.3	0.155	0.101	0.237	0.232	0.149	0.318
		0.5	0.369	0.193	0.535	0.495	0.285	0.6
		0.8	0.823	0.523	0.778	0.888	0.595	0.818
1000-1000	1/8	0.3	0.333	0.187	0.38	0.36	0.243	0.41
		0.5	0.8	0.523	0.673	0.84	0.613	0.703
		0.8	1	0.927	0.947	1	0.943	0.95
	1/4	0.3	0.32	0.173	0.437	0.373	0.21	0.478
		0.5	0.758	0.42	0.755	0.807	0.51	0.773
		0.8	0.997	0.888	0.95	0.998	0.918	0.957
	1/2	0.3	0.333	0.171	0.442	0.383	0.2	0.47
		0.5	0.777	0.425	0.737	0.816	0.488	0.753
		0.8	0.997	0.861	0.963	0.998	0.878	0.973

Table 2.16. *Power Rates across the DIF Conditions with Unequal Base Rate and Unequal Sample Size Under Short Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	1/8	0.3	0.137	0.057	0.103	0.22	0.15	0.213
		0.5	0.28	0.123	0.37	0.427	0.267	0.5
		0.8	0.677	0.273	0.727	0.843	0.453	0.81
	1/4	0.3	0.108	0.075	0.137	0.223	0.162	0.26
		0.5	0.243	0.118	0.41	0.42	0.243	0.545
		0.8	0.635	0.263	0.73	0.805	0.447	0.83
	1/2	0.3	0.122	0.073	0.143	0.232	0.155	0.273
		0.5	0.281	0.133	0.404	0.423	0.268	0.545
		0.8	0.66	0.272	0.723	0.81	0.451	0.829
1500-500	1/8	0.3	0.267	0.123	0.367	0.323	0.17	0.42
		0.5	0.66	0.253	0.677	0.72	0.367	0.717
		0.8	0.983	0.723	0.977	0.99	0.813	0.98
	1/4	0.3	0.255	0.108	0.393	0.325	0.163	0.457
		0.5	0.653	0.243	0.74	0.72	0.343	0.788
		0.8	0.977	0.653	0.973	0.985	0.737	0.98
	1/2	0.3	0.27	0.12	0.377	0.335	0.169	0.433
		0.5	0.645	0.26	0.728	0.713	0.335	0.759
		0.8	0.973	0.599	0.961	0.986	0.683	0.96

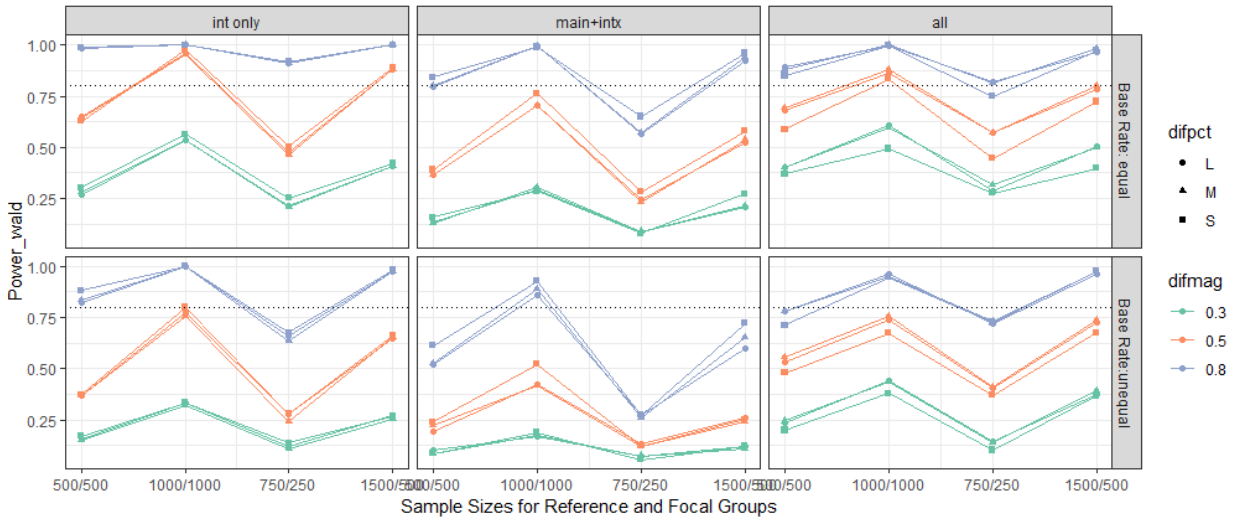

Figure 2.1. Power Rates for The Wald Method Under Short Tests

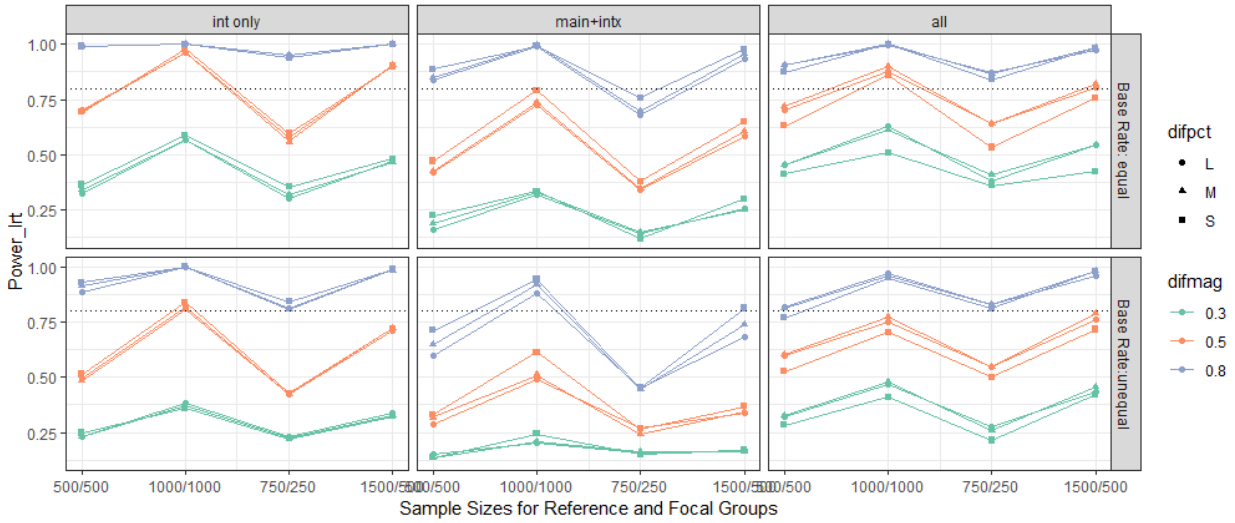

Figure 2.2. Power Rates for The LRT Method Under Short Tests

Table 2.17. *Power Rates across the DIF Conditions with Equal Base Rate and Equal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
500-500	1/8	0.3	0.228	0.115	0.317	0.4	0.237	0.423
		0.5	0.737	0.427	0.535	0.853	0.587	0.658
		0.8	0.997	0.903	0.89	1	0.962	0.962
	1/4	0.3	0.223	0.098	0.373	0.39	0.222	0.512
		0.5	0.695	0.363	0.657	0.834	0.549	0.746
		0.8	0.995	0.85	0.915	1	0.933	0.96
	1/2	0.3	0.221	0.101	0.383	0.383	0.213	0.513
		0.5	0.687	0.362	0.654	0.826	0.544	0.743
		0.8	0.996	0.839	0.913	1	0.932	0.965
1000-1000	1/8	0.3	0.612	0.337	0.547	0.697	0.41	0.603
		0.5	0.982	0.828	0.875	0.987	0.87	0.9
		0.8	1	1	0.995	1	1	0.995
	1/4	0.3	0.574	0.3	0.639	0.657	0.386	0.683
		0.5	0.975	0.8	0.897	0.986	0.855	0.923
		0.8	1	1	0.998	1	1	0.998
	1/2	0.3	0.585	0.302	0.648	0.668	0.391	0.686
		0.5	0.977	0.782	0.893	0.988	0.838	0.92
		0.8	1	0.997	0.998	1	1	0.998

Table 2.18. *Power Rates across the DIF Conditions with Equal Base Rate and Unequal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	1/8	0.3	0.073	0.035	0.127	0.337	0.217	0.397
		0.5	0.31	0.098	0.365	0.715	0.43	0.59
		0.8	0.893	0.465	0.625	0.99	0.858	0.903
	1/4	0.3	0.059	0.023	0.144	0.299	0.197	0.433
		0.5	0.291	0.082	0.463	0.695	0.414	0.698
		0.8	0.866	0.411	0.718	0.988	0.839	0.916
	1/2	0.3	0.057	0.024	0.152	0.299	0.19	0.425
		0.5	0.303	0.081	0.474	0.697	0.423	0.705
		0.8	0.852	0.402	0.712	0.983	0.82	0.919
1500-500	1/8	0.3	0.408	0.17	0.417	0.563	0.323	0.545
		0.5	0.91	0.618	0.688	0.95	0.79	0.815
		0.8	1	0.995	0.982	1	0.998	0.995
	1/4	0.3	0.393	0.147	0.507	0.533	0.284	0.637
		0.5	0.899	0.55	0.771	0.954	0.733	0.86
		0.8	0.999	0.972	0.989	0.999	0.988	0.995
	1/2	0.3	0.404	0.146	0.517	0.55	0.29	0.631
		0.5	0.898	0.55	0.78	0.948	0.735	0.865
		0.8	1	0.964	0.984	1	0.99	0.995

Table 2.19. *Power Rates across the DIF Conditions with Unequal Base Rate and Equal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald Int Only	Wald Main+inx	Wald All	LRT Int Only	LRT Main+inx	LRT All
500-500	1/8	0.3	0.063	0.033	0.18	0.318	0.18	0.34
		0.5	0.315	0.16	0.433	0.688	0.453	0.568
		0.8	0.868	0.555	0.703	0.978	0.852	0.858
	1/4	0.3	0.06	0.024	0.188	0.288	0.146	0.404
		0.5	0.283	0.123	0.518	0.663	0.396	0.669
		0.8	0.829	0.487	0.767	0.972	0.807	0.902
	1/2	0.3	0.055	0.033	0.186	0.286	0.155	0.405
		0.5	0.291	0.131	0.505	0.655	0.39	0.672
		0.8	0.819	0.462	0.765	0.974	0.795	0.896
1000-1000	1/8	0.3	0.37	0.158	0.415	0.533	0.298	0.49
		0.5	0.87	0.57	0.712	0.933	0.725	0.772
		0.8	1	0.972	0.977	1	0.992	0.988
	1/4	0.3	0.32	0.158	0.497	0.491	0.265	0.578
		0.5	0.843	0.515	0.783	0.922	0.686	0.831
		0.8	1	0.948	0.979	1	0.983	0.992
	1/2	0.3	0.325	0.161	0.505	0.494	0.277	0.591
		0.5	0.848	0.491	0.781	0.928	0.663	0.836
		0.8	1	0.93	0.977	1	0.973	0.993

Table 2.20. *Power Rates across the DIF Conditions with Unequal Base Rate and Unequal Sample Size Under Long Tests*

Sample Size (R-F)	DIF Percentage	DIF Size	Wald			LRT		
			Int Only	Main+inx	All	Int Only	Main+inx	All
750-250	1/8	0.3	0.027	0	0.032	0.253	0.135	0.308
		0.5	0.12	0.005	0.203	0.562	0.272	0.565
		0.8	0.495	0.04	0.612	0.937	0.633	0.883
	1/4	0.3	0.019	0.003	0.028	0.247	0.136	0.315
		0.5	0.103	0.008	0.196	0.548	0.24	0.637
		0.8	0.449	0.034	0.633	0.917	0.593	0.903
	1/2	0.3	0.025	0.002	0.03	0.24	0.133	0.308
		0.5	0.107	0.008	0.206	0.563	0.257	0.654
		0.8	0.442	0.038	0.624	0.92	0.563	0.9
1500-500	1/8	0.3	0.223	0.065	0.313	0.428	0.18	0.477
		0.5	0.705	0.21	0.665	0.86	0.502	0.76
		0.8	0.995	0.693	0.972	1	0.932	0.99
	1/4	0.3	0.205	0.047	0.333	0.42	0.168	0.529
		0.5	0.673	0.16	0.737	0.842	0.442	0.825
		0.8	0.99	0.593	0.985	0.999	0.861	0.993
	1/2	0.3	0.221	0.051	0.34	0.425	0.175	0.524
		0.5	0.678	0.175	0.741	0.848	0.446	0.827
		0.8	0.99	0.57	0.976	0.999	0.844	0.991

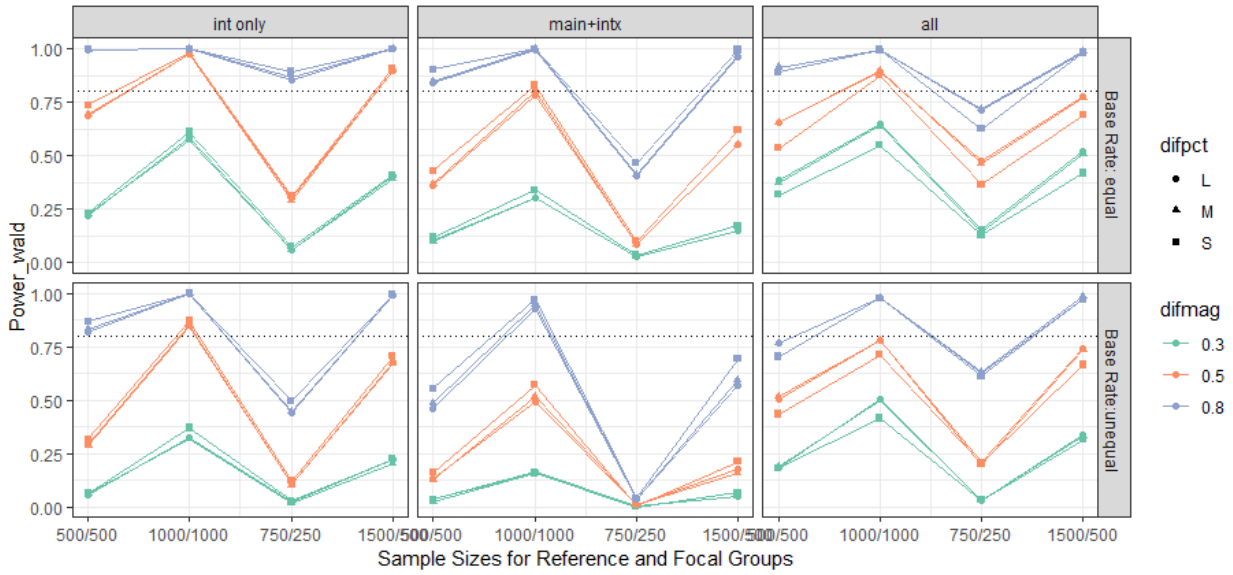

Figure 2.3. Power Rates for The Wald Method Under Long Tests

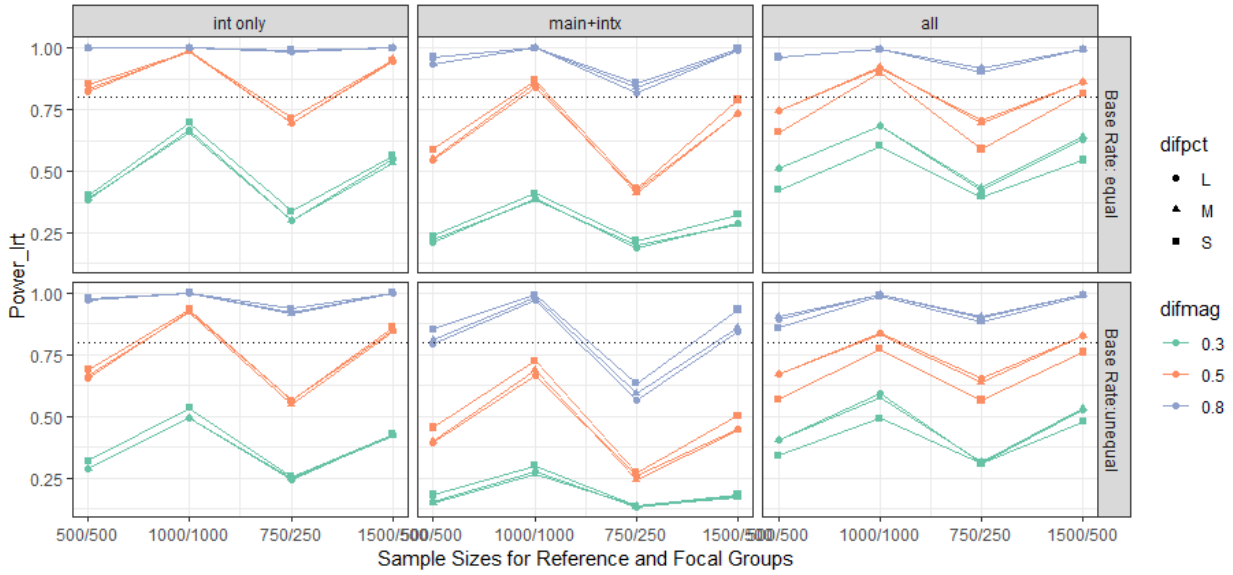

Figure 2.4. Power Rates for The LRT Method Under Long Tests

Table 3.1. *Q-matrix for the Simulation Study II*

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	1	0	0
6	1	0	1	0
7	0	1	0	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0
11	0	1	1	0
12	0	1	0	1
13	0	0	1	0
14	0	0	1	0
15	0	0	1	0
16	0	0	1	0
17	0	0	1	1
18	1	0	1	0
19	0	0	0	1
20	0	0	0	1
21	0	0	0	1
22	0	0	0	1
23	1	0	0	1
24	0	1	0	1

Table 3.2. *Summary of Simulation Conditions for Simulation Study II*

Manipulated Factors		Value	
Sample Sizes	Equal	$N_R = 1000$, $N_F = 1000$	
	Unequal	$N_R = 1500$, $N_F = 500$	
Percentages of DIF items		0, 1/8, 1/4, 1/2 (0, 3, 6, 12 items)	
DIF Magnitude		$\Delta_{\lambda_{i,0}} = 0 \pm .1k$, $\Delta_{\lambda_i} = 0 \pm .1k$, $k \in \{0,1,2,\ldots,9,10\}$	
DIF Type		$\Delta_{\lambda_{i,0}} = (\lambda_{i,0})_F - (\lambda_{i,0})_R$	$\Delta_{\lambda_i} = (\lambda_i)_F - (\lambda_i)_R$
	Intercept only	$-$	0
	Main effect and interactions	0	$-$
	All parameters	$-$	$-$

Note. N_R = Reference group sample size; N_F = Focal group sample size; $-$ indicates a negative number, 0 indicates no difference between groups; $\Delta_{\lambda_{i,0}}$ = The difference in the intercept parameters between the focal and reference groups, $\Delta_{\boldsymbol{\lambda_i}}$ = The difference in the main effects and possible interaction terms for item *i* between the focal and reference groups.

Table 3.3. *Combination of DIF Sizes and DIF Types for Simulation Study II*

DIF Type	DIF Size	$\Delta_{\boldsymbol{\lambda_{i,0}}} = (\boldsymbol{\lambda_{i,0}})_{\boldsymbol{F}} - (\boldsymbol{\lambda_{i,0}})_{\boldsymbol{R}}$	$\Delta_{\boldsymbol{\lambda_i}} = (\boldsymbol{\lambda_i})_{\boldsymbol{F}} - (\boldsymbol{\lambda_i})_{\boldsymbol{R}}$
No DIF	0	0	0
Intercept only	.1	$-.1$	0
	.2	$-.2$	0
	…	…	…
	.9	$-.9$	0
	1	-1	0
Main effect and interactions	.1	0	$-.1$
	.2	0	$-.2$
	...	…	…
	.9	0	$-.9$
	1	0	-1
All parameters	.1	$-.1$	$-.1$
	.2	$-.2$	$-.2$
	...	…	…
	.9	$-.9$	$-.9$
	1	-1	-1

Note. $\Delta_{\lambda_{i,0}}$ = The difference in the intercept parameters between the focal and reference groups, $\Delta_{\boldsymbol{\lambda_i}}$ = The difference in the main effects and possible interaction terms for item *i* between the focal and reference groups.

Table 3.4. *Classification Accuracy Rates for Focal Group (PCA and ACA), $N_R = 1000$, $N_F = 1000$*

			PCA			ACA		
Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All	Int Only	Main+inx	All
1000-1000	0	0	0.649	0.649	0.649	0.899	0.899	0.899
	1/8	0.1	0.65	0.647	0.647	0.899	0.898	0.898
		0.2	0.649	0.643	0.643	0.899	0.896	0.896
		0.3	0.65	0.639	0..639	0.899	0.895	0.895
		0.4	0.649	0.635	0.635	0.898	0.893	0.894
		0.5	0.646	0.634	0.633	0.897	0.893	0.893
		0.6	0.647	0.632	0.632	0.898	0.893	0.892
		0.7	0.643	0.628	0.629	0.897	0.891	0.891
		0.8	0.642	0.624	0.627	0.896	0.890	0.891
		0.9	0.641	0.625	0.622	0.896	0.890	0.889
		1	0.639	0.618	0.623	0.895	0.888	0.889
	1/4	0.1	0.651	0.643	0.643	0.899	0.896	0.896
		0.2	0.649	0.637	0.637	0.899	0.894	0.894
		0.3	0.652	0.633	0.629	0.899	0.893	0.891
		0.4	0.647	0.625	0.626	0.897	0.890	0.890
		0.5	0.644	0.619	0.616	0.897	0.888	0.887
		0.6	0.642	0.615	0.609	0.895	0.887	0.884
		0.7	0.637	0.611	0.603	0.895	0.885	0.882
		0.8	0.634	0.605	0.597	0.894	0.883	0.880
		0.9	0.632	0.597	0.590	0.893	0.881	0.877
		1	0.628	0.593	0.586	0.891	0.879	0.877
	1/2	0.1	0.650	0.637	0.637	0.898	0.894	0.894
		0.2	0.645	0.623	0.625	0.897	0.890	0.890
		0.3	0.646	0.613	0.609	0.898	0.886	0.884
		0.4	0.642	0.601	0.595	0.896	0.882	0.880
		0.5	0.636	0.587	0.583	0.894	0.877	0.875
		0.6	0.627	0.574	0.568	0.891	0.872	0.869
		0.7	0.621	0.561	0.554	0.889	0.867	0.864
		0.8	0.613	0.551	0.538	0.886	0.863	0.858
		0.9	0.607	0.543	0.527	0.884	0.860	0.853
		1	0.597	0.529	0.513	0.880	0.855	0.848

Table 3.5. *Classification Accuracy Rates for Focal Group (PCA and ACA), $N_R = 1500$, $N_F = 500$*

			PCA			ACA		
Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All	Int Only	Main+inx	All
1500-500	0	0	0.650	0.650	0.650	0.899	0.899	0.899
	1/8	0.1	0.649	0.646	0.649	0.898	0.897	0.898
		0.2	0.651	0.644	0.642	0.899	0.896	0.895
		0.3	0.647	0.643	0.639	0.897	0.897	0.895
		0.4	0.646	0.642	0.637	0.897	0.896	0.894
		0.5	0.645	0.635	0.632	0.897	0.893	0.892
		0.6	0.644	0.630	0.629	0.896	0.891	0.891
		0.7	0.641	0.624	0.624	0.896	0.890	0.890
		0.8	0.640	0.625	0.622	0.895	0.890	0.889
		0.9	0.639	0.618	0.620	0.895	0.888	0.888
		1	0.639	0.621	0.618	0.894	0.889	0.888
	1/4	0.1	0.649	0.641	0.640	0.898	0.896	0.896
		0.2	0.651	0.638	0.640	0.898	0.894	0.895
		0.3	0.649	0.632	0.632	0.898	0.892	0.892
		0.4	0.648	0.624	0.622	0.898	0.889	0.889
		0.5	0.643	0.619	0.615	0.897	0.888	0.886
		0.6	0.641	0.605	0.608	0.895	0.884	0.884
		0.7	0.638	0.607	0.597	0.895	0.884	0.880
		0.8	0.632	0.597	0.591	0.893	0.881	0.878
		0.9	0.624	0.591	0.582	0.890	0.878	0.875
		1	0.620	0.581	0.577	0.888	0.875	0.873
	1/2	0.1	0.652	0.635	0.640	0.899	0.894	0.895
		0.2	0.651	0.627	0.622	0.899	0.891	0.889
		0.3	0.642	0.613	0.606	0.896	0.886	0.884
		0.4	0.636	0.595	0.588	0.894	0.880	0.878
		0.5	0.632	0.585	0.576	0.892	0.876	0.873
		0.6	0.625	0.567	0.558	0.890	0.870	0.866
		0.7	0.625	0.553	0.544	0.886	0.864	0.860
		0.8	0.600	0.534	0.521	0.882	0.857	0.852
		0.9	0.587	0.521	0.498	0.876	0852	0.843
		1	0.573	0.511	0.490	0.871	0.847	0.840

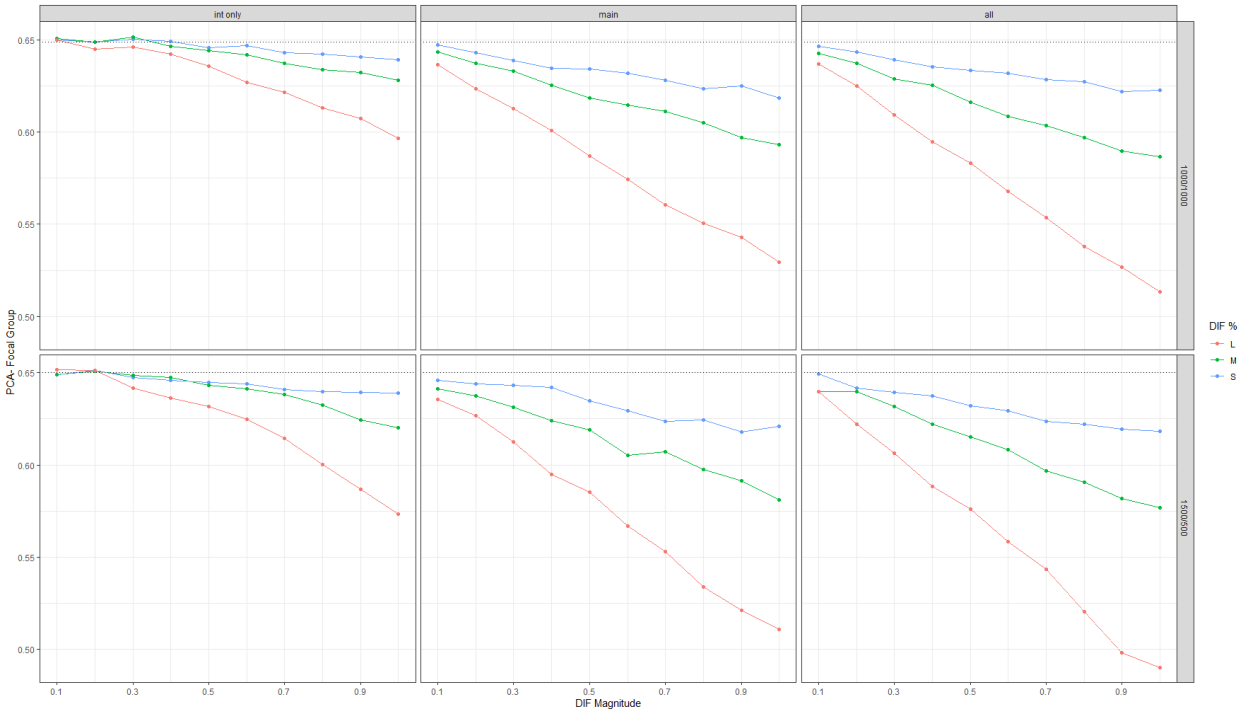

Figure 3.1. Profile Level Classification Accuracy Rates for Focal Group. Note. The dotted lines represent the classification accuracies for baseline conditions with no DIF.

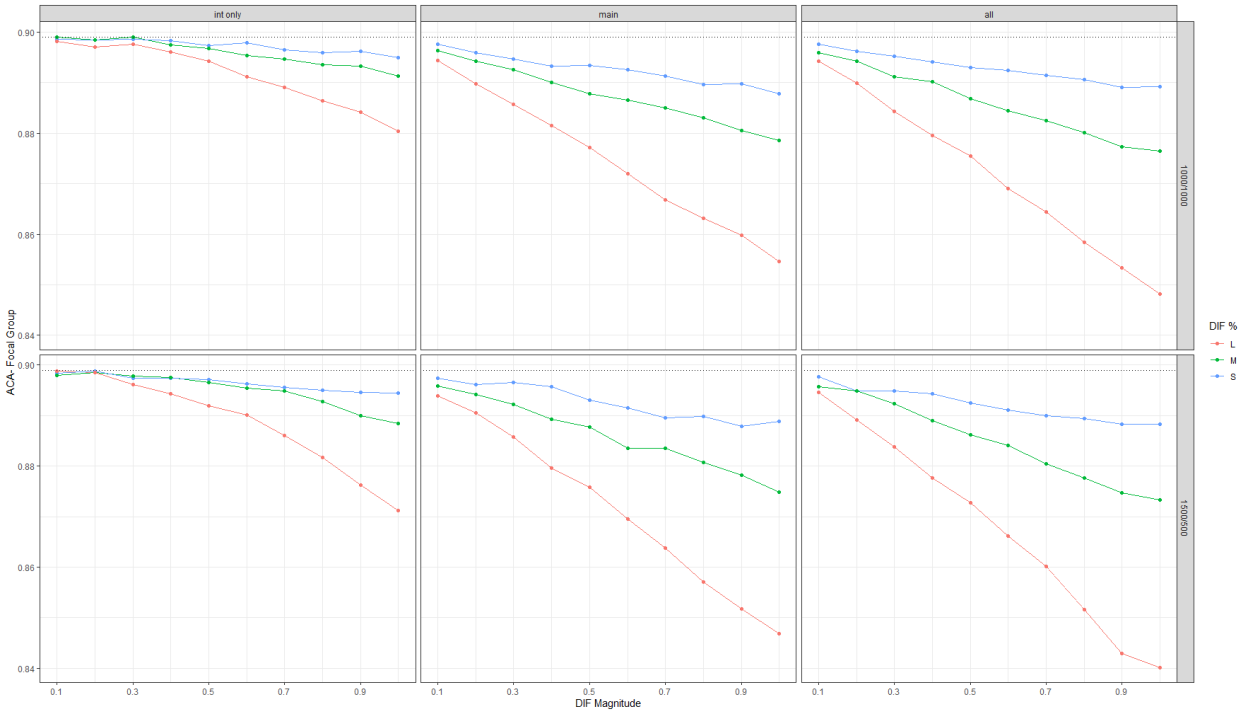

Figure 3.2. Attribute Level Classification Accuracy Rates for Focal Group. Note. The dotted lines represent the classification accuracies for baseline conditions with no DIF.

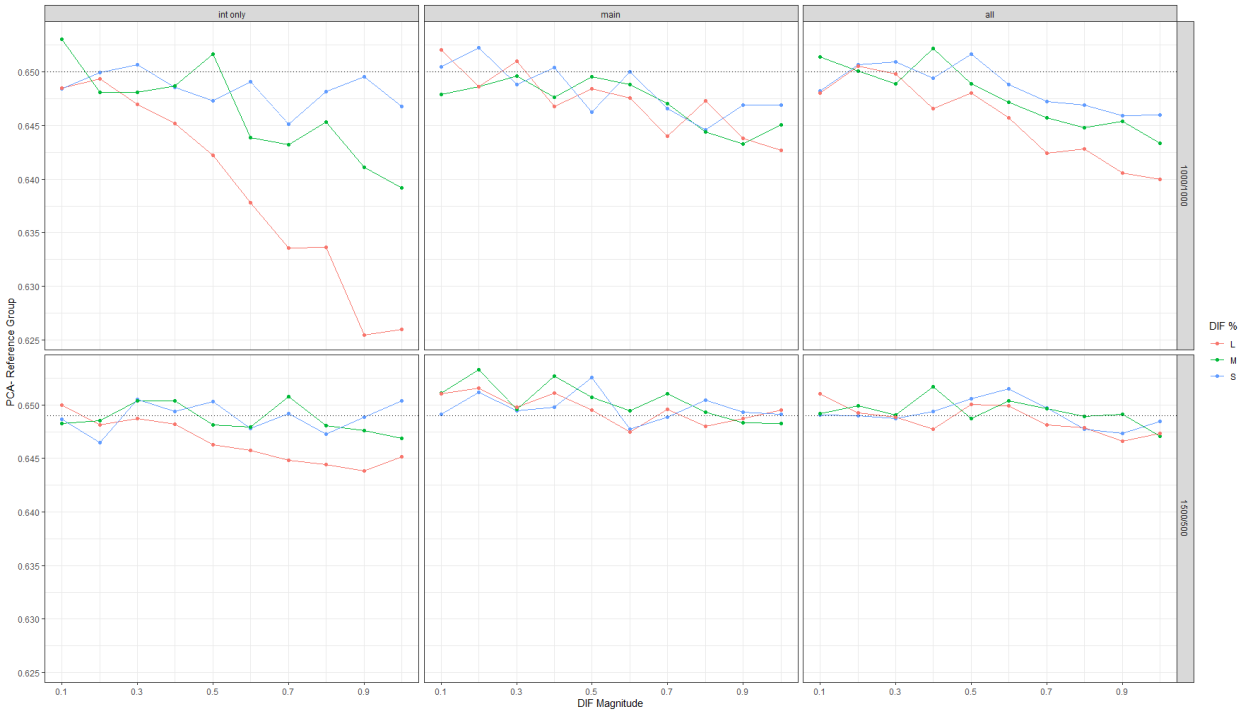

Figure 3.3. Profile Level Classification Accuracy Rates for Reference Group

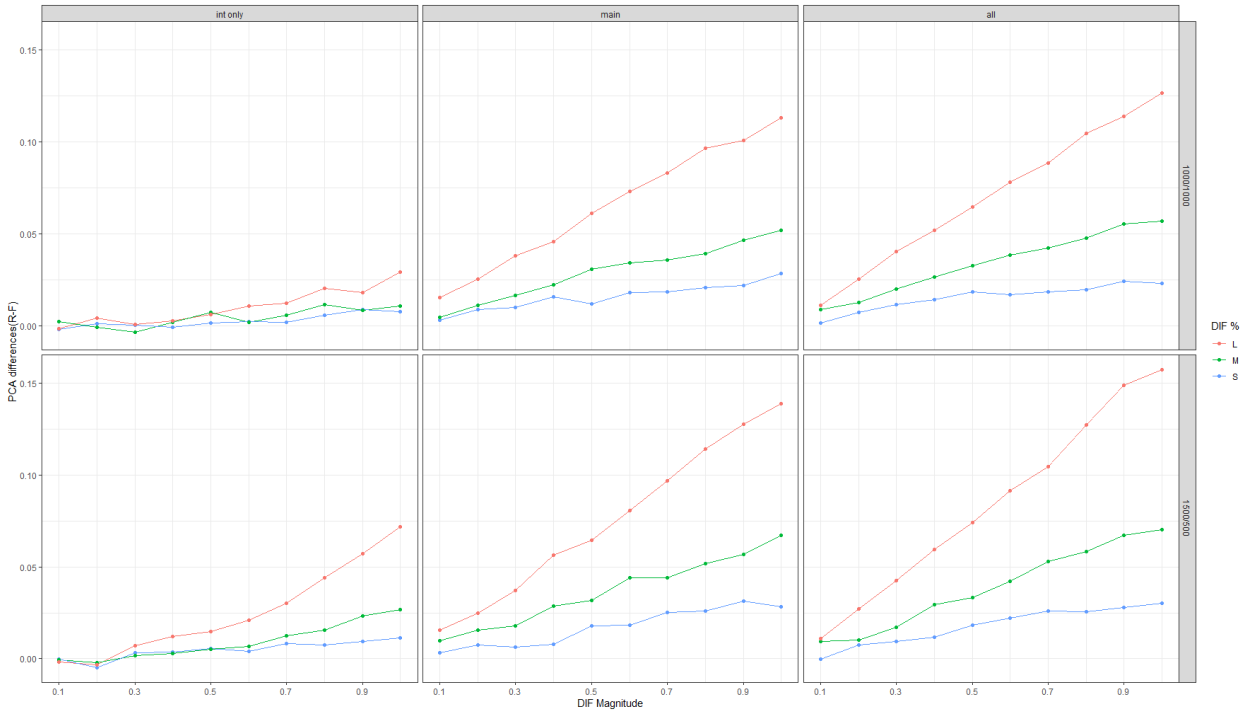

Figure 3.4. Differences in Profile level Classification Accuracy Rates for Focal and Reference Groups

Table 3.6. *Classification Reliabilities Under the Equal Sample Sizes ($N_R = 1000$, $N_F = 1000$)*

Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All
1000-1000	0	0	0.803	0.803	0.803
	1/8	0.1	0.804	0.802	0.801
		0.2	0.804	0.802	0.801
		0.3	0.804	0.799	0.8
		0.4	0.804	0.798	0.799
		0.5	0.802	0.796	0.796
		0.6	0.802	0.796	0.796
		0.7	0.801	0.792	0.792
		0.8	0.8	0.789	0.79
		0.9	0.799	0.79	0.788
		1	0.798	0.788	0.788
	1/4	0.1	0.804	0.801	0.8
		0.2	0.803	0.797	0.798
		0.3	0.804	0.794	0.795
		0.4	0.803	0.793	0.792
		0.5	0.802	0.79	0.788
		0.6	0.8	0.785	0.783
		0.7	0.799	0.784	0.778
		0.8	0.797	0.78	0.778
		0.9	0.796	0.777	0.772
		1	0.794	0.774	0.769
	1/2	0.1	0.803	0.798	0.798
		0.2	0.802	0.793	0.793
		0.3	0.802	0.786	0.784
		0.4	0.799	0.778	0.776
		0.5	0.799	0.773	0.77
		0.6	0.797	0.767	0.76
		0.7	0.793	0.76	0.754
		0.8	0.792	0.754	0.745
		0.9	0.787	0.749	0.738
		1	0.783	0.743	0.731

Table 3.7. *Classification Reliabilities Under the Unequal Sample Sizes ($N_R = 1500$, $N_F = 500$)*

Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All
1500-500	0	0	0.804	0.804	0.804
	1/8	0.1	0.804	0.803	0.805
		0.2	0.803	0.804	0.801
		0.3	0.803	0.801	0.802
		0.4	0.804	0.801	0.801
		0.5	0.803	0.8	0.799
		0.6	0.802	0.799	0.798
		0.7	0.802	0.797	0.797
		0.8	0.801	0.796	0.797
		0.9	0.802	0.797	0.795
		1	0.8	0.795	0.795
	1/4	0.1	0.804	0.802	0.802
		0.2	0.804	0.801	0.801
		0.3	0.802	0.8	0.8
		0.4	0.802	0.798	0.797
		0.5	0.803	0.796	0.794
		0.6	0.801	0.794	0.792
		0.7	0.8	0.793	0.791
		0.8	0.801	0.791	0.789
		0.9	0.8	0.789	0.787
		1	0.799	0.788	0.784
	1/2	0.1	0.804	0.801	0.8
		0.2	0.804	0.797	0.798
		0.3	0.803	0.795	0.794
		0.4	0.802	0.793	0.789
		0.5	0.8	0.789	0.786
		0.6	0.8	0.785	0.781
		0.7	0.799	0.782	0.778
		0.8	0.796	0.777	0.773
		0.9	0.795	0.775	0.768
		1	0.793	0.774	0.767

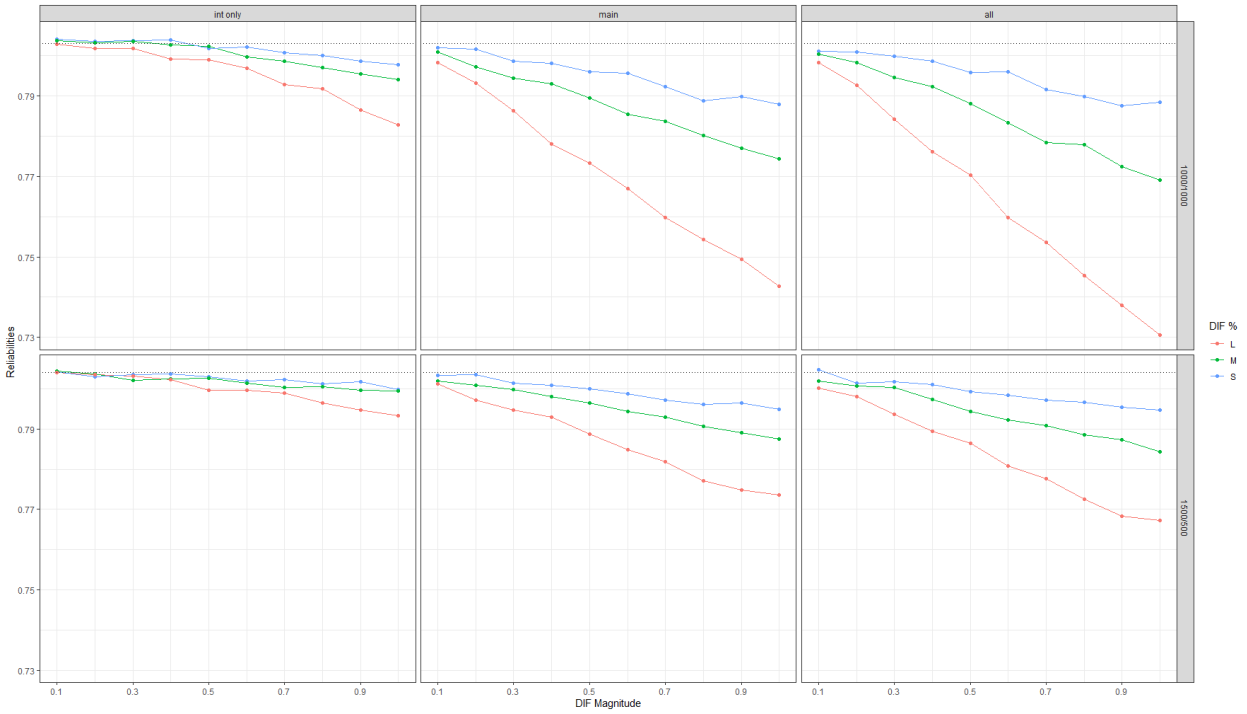

Figure 3.5. Classification Reliabilities Across the DIF conditions

Table 3.8. *Under Area (UA) Effect Sizes Under the Equal Sample Sizes ($N_R = 1000$, $N_F = 1000$)*

Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All
1000-1000	0	0	0.026	0.026	0.026
	1/8	0.1	0.027	0.027	0.027
		0.2	0.029	0.028	0.029
		0.3	0.031	0.029	0.031
		0.4	0.034	0.03	0.033
		0.5	0.036	0.032	0.036
		0.6	0.039	0.034	0.038
		0.7	0.042	0.035	0.039
		0.8	0.044	0.037	0.042
		0.9	0.047	0.038	0.045
		1	0.049	0.04	0.046
	1/4	0.1	0.028	0.027	0.028
		0.2	0.03	0.029	0.033
		0.3	0.035	0.031	0.038
		0.4	0.039	0.034	0.045
		0.5	0.044	0.037	0.05
		0.6	0.051	0.039	0.056
		0.7	0.056	0.043	0.062
		0.8	0.062	0.046	0.068
		0.9	0.067	0.049	0.073
		1	0.072	0.053	0.077
	1/2	0.1	0.028	0.028	0.031
		0.2	0.033	0.03	0.039
		0.3	0.041	0.034	0.05
		0.4	0.049	0.039	0.062
		0.5	0.058	0.045	0.075
		0.6	0.067	0.051	0.088
		0.7	0.077	0.057	0.099
		0.8	0.087	0.063	0.11
		0.9	0.096	0.069	0.123
		1	0.105	0.075	0.137

Table 3.9. *Under Area (UA) Effect Sizes Under the Unequal Sample Sizes ($N_R = 1500$, $N_F = 500$)*

Sample Size (R-F)	DIF Percentage	DIF Size	Int Only	Main+inx	All
1500-500	0	0	0.03	0.03	0.03
	1/8	0.1	0.032	0.031	0.031
		0.2	0.033	0.032	0.033
		0.3	0.035	0.033	0.035
		0.4	0.037	0.035	0.037
		0.5	0.04	0.036	0.039
		0.6	0.042	0.037	0.042
		0.7	0.045	0.039	0.044
		0.8	0.047	0.04	0.046
		0.9	0.05	0.042	0.049
		1	0.053	0.043	0.051
	1/4	0.1	0.031	0.031	0.033
		0.2	0.034	0.032	0.036
		0.3	0.038	0.034	0.042
		0.4	0.043	0.037	0.048
		0.5	0.048	0.04	0.054
		0.6	0.053	0.044	0.06
		0.7	0.057	0.047	0.066
		0.8	0.064	0.05	0.071
		0.9	0.069	0.053	0.077
		1	0.074	0.055	0.083
	1/2	0.1	0.032	0.031	0.034
		0.2	0.036	0.034	0.042
		0.3	0.042	0.038	0.053
		0.4	0.051	0.042	0.064
		0.5	0.059	0.046	0.078
		0.6	0.069	0.054	0.091
		0.7	0.079	0.059	0.103
		0.8	0.088	0.065	0.113
		0.9	0.099	0.072	0.126
		1	0.107	0.078	0.14

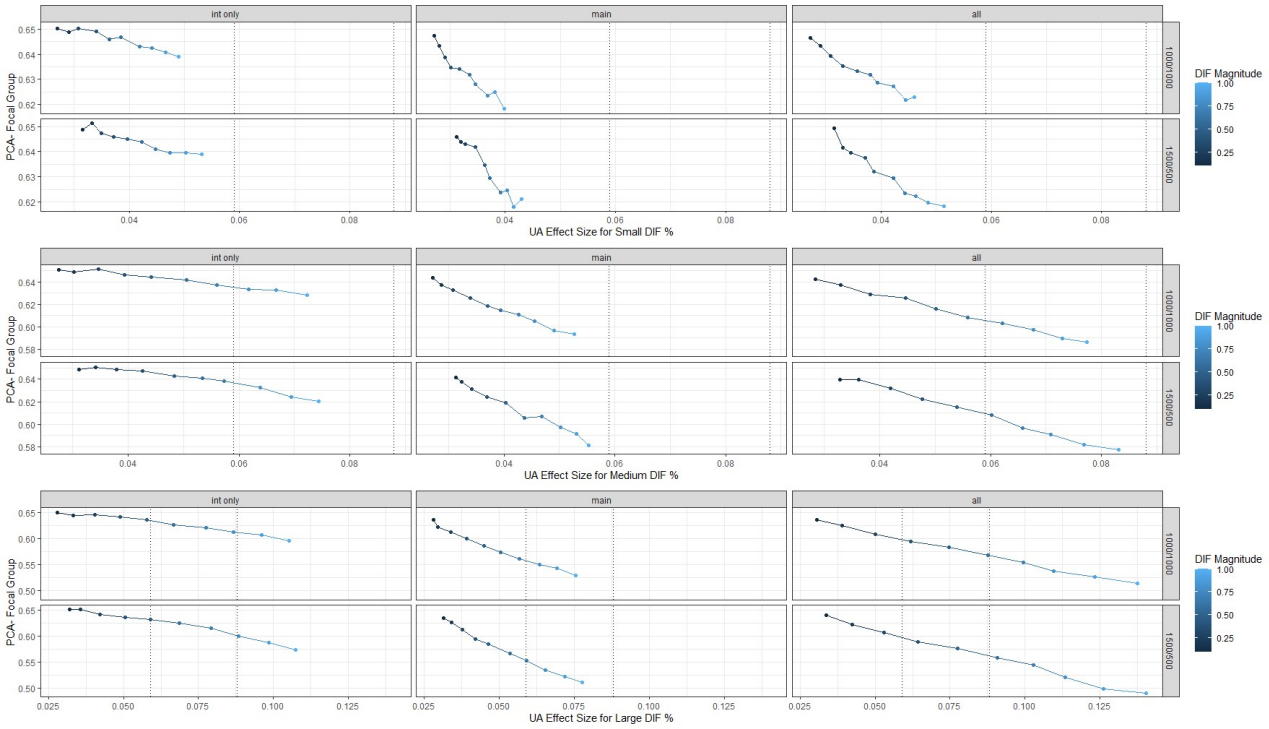

Figure 3.6. Profile Classification Accuracies versus the UA effect size estimates