2020

Assessing the Performance of Two Procedures for Detecting Differential Item Functioning within the Multilevel Partial Credit Model

Carol Hanley
University of Kentucky, chanley@uky.edu
Author ORCID Identifier:
https://orcid.org/0000-0002-9665-4599
Digital Object Identifier: https://doi.org/10.13023/etd.2020.073

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Carol Hanley, Student

Dr. Michael D. Toland, Major Professor

Dr. Michael D. Toland, Director of Graduate Studies

ASSESSING THE PERFORMANCE OF TWO PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITHIN THE MULTILEVEL PARTIAL CREDIT MODEL

DISSERTATION

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy in the

College of Education

at the University of Kentucky

By

Carol D. Hanley

Lexington, Kentucky

Co-Directors: Dr. Michael D. Toland, Professor of Quantitative and Psychometric Methods

and Dr. Xin Ma, Professor of Quantitative and Psychometric Methods

Lexington, Kentucky

https://orcid.org/0000-0002-9665-4599

ABSTRACT OF DISSERTATION

ASSESSING THE PERFORMANCE OF TWO PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITHIN THE MULTILEVEL PARTIAL CREDIT MODEL

This is a simulation study that evaluates the performances of two models for the detection of uniform differential item functioning (DIF). Simulated data are generated by a multilevel partial credit model (MLPCM). The purpose of this study was to compare the accuracy of two DIF detection procedures, hierarchical ordinal logistic regression (HOLR) for multilevel data and multilevel generalized Mantel-Haenszel (MGMH: French & Finch, 2013; French, Finch, & Imekus, 2019). Conditions manipulated were the number of participants per cluster (20, 40), number of clusters (50, 100, 200), DIF magnitude (0, .4, .8), and magnitude of intraclass correlation coefficient (.05, .25, .45). Furthermore, only one grouping variable was used within-groups. Data was simulated using R (R Core Team, 2019), whereas analyses will be performed using SAS 9.4 (SAS Institute, 2013) and R. In general, HOLR maintains the Type I error rate better than MGMH and HOLR has more power than MGMH under most simulation conditions.

Keywords: item response theory, multilevel partial credit model, hierarchical ordinal logistic regression, multilevel generalized Mantel-Haenszel, multilevel differential item functioning

Carol D. Hanley

April 14, 2020

ASSESSING THE PERFORMANCE OF TWO PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITHIN THE MULTILEVEL PARTIAL CREDIT MODEL

By

Carol D. Hanley

Michael D. Toland, PhD

Co-Director of Dissertation

Xin Ma, PhD

Co-Director of Dissertation

Michael D. Toland, PhD

Director of Graduate Studies

April 7, 2020

ACKNOWLEDGEMENTS

In 1994, Jane Cowen-Fletcher, a Peace Corps volunteer in Benin, West Africa, wrote a book called "It takes a village." The story line begins as a mother asks her daughter, Yemi, to watch her younger brother. Yemi proudly declares that she can watch him "All by myself." Yemi loses sight of her brother and soon discovers she is not doing the task, "All by myself," but that her brother is being cared for by many villagers. Once reunited, Yemi thanks each villager in turn and comes to understand what it means to belong to a community.

As with Yemi, I could not complete this dissertation "All by myself" but had many villagers playing different roles. Some villagers were teachers, while others were tutors, supporters, and cheerleaders. The best wore multiple hats. No matter the role, they all must be thanked in turn.

My faculty advisors gave me their most precious resource – time. This is a resource that should never be undervalued or squandered and one that I highly treasure. Drs. Schroeder and Hammer supported me through my proposal defense and dissertation defense, and I thank them for sharing their time and expertise with me. Dr. Ma taught and mentored me from start to finish, and I thank him for his patience, self-lessness, humor, ability to simplify complex concepts, and make connections among statistical concepts. Dr. Toland never failed to ask a question that was just beyond my grasp of understanding, making me a better student and researcher. Thank you for your enthusiasm, expertise, dedication, and insightfulness.

Drs. Ma and Toland were my village mayors. Sometimes, they tried to teach me things I was not ready to learn, and when I was ready to learn them, I had to teach myself.

Somehow, they made me reach deeper than I wanted to, but they knew I could, and they gave me the tools to do it. That was a real gift.

A most important villager, who acted alternately as a colleague, tutor, and illustrator, was David – a savant-like friend, if I may call him that, who helped me find my way out of many intellectual bewilderments that seemed unassailable.

One villager played the role of IT specialist, not someone who was present in Yemi's community, but she would have benefited greatly. Chris was always available to install or update just one more software program to help me on my journey. He was unfailingly supportive, and I appreciate him greatly. XOXO

But villagers do not only play academic roles, they also provide emotional support. For that support, I want to thank two good friends, Melody and Jackie, and my sister, Christie. Whether on long walks or long phone calls, I was sure to have someone who would listen to a complaint – or two!

Yemi's community did not have nurses but mine does, and I am so thankful. Deborah, Grace, Janke, and Lisa were behind me to push, pull, and catch me when things didn't go my way.

My lab mates in the Applied Psychometric Strategies Lab offered a sense of community as I traveled the path to my degree. I big thank you to them all.

Some of my greatest thanks goes to a villager with superhuman listening, empathetic, and supportive powers – Esther. She never failed to ask how things were going, give me encouragement, and find the positive in very dismal outcomes, of which there were many. The story of Esther has historically been one of honor and friendship and that story continues today. Thank you, friend.

Sometimes, I had to communicate with an adjoining village, which had resources that were absent from mine. This village, the Applied Statistics Lab in University's College of Arts & Sciences, played a life-saving role. Three consultants, Eric, Matt, and Leon assisted me while I learned R coding, and I send a hearty thank you. The mayor of this village, Dr. Arnold Stromberg, deserves special recognition because of how he understands and carries out his role at the University. Unfortunately, there is a dearth of people like him in this world, and we are poorer because of it.

I would like to conclude my village tour by thanking the village founders and paraphrasing a useful quote. If I have been able to investigate this subject a bit more deeply, it is because I have been standing upon the shoulders of those who have who have conducted foundational work. I sincerely thank Dr. Brian French and Dr, Holmes Finch for making my work possible.

TABLE OF CONTENTS

LIST OF TABLES

Chapter One: Introduction

The Importance of Differential Item Functioning

In our current political and educational environment, policymakers and educators rely heavily on standardized test scores to inform their decisions regarding schools, teachers, and students. To ensure these decisions are equitable, test developers must create assessment items that are fair for every test taker. Fairness in this case means that different groups of test takers, who have the same abilities, should have the same probability of getting any item correct, and test scores must accurately reflect each test taker's ability on the construct of interest (e.g., reading achievement). The federal law, Every Student Succeeds Act 2015 (ESSA, 2015), requires states to use assessments that "are valid, reliable, and comparable for all students and for each subgroup of students and among participating schools and districts" (U.S. Department of Education, 2017, p. 5). For these reasons, accurate methods must exist to determine if items or tests contain bias.

Test bias is defined as the systematic error in how a test measures members of a particular group and creates a distortion in results for one group over another (Camilli & Shepard, 1994). Differential item functioning (DIF), on the other hand, can be defined as "an unexpected difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and the test on which it appears" (Dorans & Holland, 1993, p. 37). Camilli and Shepard compared DIF to bias, asserting that DIF is an item's psychometric property, whereas bias is a general term associated with interpretation. Bias can be defined as "construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA et al., 2003, p. F6). In the high-stakes environment of state assessments, these

construct-irrelevant components may include gender, race, ethnicity, socioeconomic status, rurality, religion, sexual identity, or any combination thereof.

DIF detection procedures were developed because of the possibility of bias in achievement tests (Wen, 2014). When examining items within a test or scale, it is essential to identify those items that contain DIF; however, it is not beneficial to flag items as containing DIF if they function as the test developer intended. Although there are many ways of considering test fairness, in this study, it will be examined through two DIF detection procedures.

DIF versus Impact

DIF should be distinguished from impact. While DIF matches students on their ability level in order to compare scores on an item, any difference in the two groups being compared in terms of overall test score means can still differ; however, this circumstance is not categorized as DIF. DIF refers to the situation in which different groups with equal ability do not have an equal chance of getting an item correct. An item is described as having DIF if it favors one group over another, such as one minority group over another while controlling for ability level. When a researcher attempts to identify items on which examinees with the same ability, but from different groups, respond differently, that researcher focuses on DIF, not on impact (Kim, 1992). When true differences in group performance exist due to proficiency, this situation is referred to as an impact. Impact is defined as the influence on the probability of correctly answering the target item based on the presence of differences on ability dimension between groups (Klokars & Lee, 2008). Therefore, when DIF is not present, but there are true differences in groups, then impact is present.

Kim (2003) suggests the differences in item performance be thought of as the "true" difference between groups and an "artificial" difference brought about by inappropriate and irrelevant (biased) items. In general, we desire for there to be no DIF among items on our test so that impact can be truly detected.

The Perils of Ignoring Data Structure

DIF detection procedures help ensure assessment items are bias-free, but they must be chosen to fit the data structure and the item response format, equally well (French & Finch, 2010, 2013). Unfortunately, DIF detection procedures are not always chosen to fit the data structure. For example, in educational environments, data are often hierarchically structured, meaning students are nested within teachers who may be further nested within schools. Although DIF detection procedures have been used for many years, often they are not conducted using multilevel methods that account for hierarchical data structure (French & Finch, 2010, 2013; Ryan, 2008; Wen, 2014). In general, when the multilevel data structure is not taken into account, it can lead to inaccurate estimation of standard errors (Raudenbush & Bryk, 2002). In regard to DIF detection, ignoring the multilevel data structure can result in parameter estimation problems, which can lead to biased statistical tests and faulty DIF detection (French & Finch).

Multilevel DIF detection procedures have mostly been used with items that are dichotomously scored (1 = correct, 0 = incorrect). However, large-scale assessments contain a mix of dichotomously and polytomously scored (0 = no credit, 1 = partial credit, 2 = full credit) items, and more polytomously scored items may be on the way. For instance, the ESSA (ESSA of 2015, 2015) gives states the opportunity to design new types of assessments to better support teaching and learning by measuring higher-order

thinking skills via more authentic assessments. These authentic assessments include projects and extended performance tasks (Holahan, Young, Palmer, & Little, 2017). As states move from traditional items, such as multiple-choice and constructed-response, that assess a single attribute, to those that assess multiple competencies (Castle, 2018), they must think about different ways to design and score assessments and detect DIF. For example, Massachusetts included multiple item formats in their 2018 Massachusetts Comprehensive Assessment System mathematics tests (Massachusetts Department of Education, 2018), which may call for multiple forms of DIF detection.

Brief Overview of Polytomous Item Response Models – Single-Level and Multilevel

Many psychological constructs are assessed by polytomous formats rather than dichotomous scoring (Preston & Reise, 2014). Whereas dichotomous items are scored in a binary way, polytomous items have more than two possible scores and describe the probability of a test taker reaching a specific score category. Thissen and Steinberg (1986) categorized polytomous response models into difference and divide-by-total models. Difference models may be used for ordered responses (i.e., options are in a specific, meaningful order) and are those in which the probability of responding in a category is found by determining the difference between cumulative probabilities. Divide-by-total models, used with either nominal data (i.e., not ordered) or ordered data, such as the partial credit model (PCM; Masters, 1982), are those in which the exponent is divided by the sum of all the exponents that appear in the numerator.

Classical item response theory single-level models include the PCM, generalized partial credit model (GPCM; Muraki, 1992), graded response model (GRM; Samejina, 1969), Andrich's (1978) rating scale model (RSM), and Bock's (1972) nominal response

model (NRM). The PCM is an ordered category response model for polytomous data that is commonly used with achievement outcomes or items that can be given partial credit; however, it can be used in any situation in which the test taker has two or more ordered category choices (Kim, 2018; Masters & Wright, 1992). Andrich's (1978) RSM can be used with Likert-type data, as well as performance data. In this model, thresholds on the latent continuum separate adjacent categories, which are constrained across items. The GPCM can be used with ordered response data, ratings, or Likert-type responses. In the GPCM, the assumption of equal item discriminations across items is relaxed, meaning the model contains a discrimination parameter that indicates the degree to which an item can differentiate among trait level values. Samejima's (1969) GRM uses a different logic from the adjacent or divide-by-total models, wherein the polytomous scores are a series of cumulative comparisons, while allowing discrimination to vary across items. Bock's (1972) NRM was originally designed for data with no order. However, the PCM, GPCM, and RSM are all special cases of the NRM; therefore, the NRM can be applied to ordinal data with specific constraints placed on model parameters. Furthermore, the NRM can be used with dichotomous data when specific constraints are placed on model parameters, which is better known as the Rasch model for dichotomous data.

Hedeker (2008) described multilevel models for categorical data that accommodate multiple random effects and allow for covariates. He viewed ordinal and nominal models as different ways of generalizing the dichotomous response model. He stated that ordinal models use cumulative dichotomizations of categorical outcomes, while nominal models use dichotomizations based on the selection of one category as a reference to which all others are compared. Unlike classical item response theory (IRT)

models, the multilevel formula of the model allows multiple covariates at either level (i.e., item-level and/or person-level covariates), which enables multilevel models to examine whether item parameters vary by personal characteristics (Hedeker).

Multilevel Data Structure Defined

Administering a standard assessment to individual test takers in different teachers' classrooms in different schools creates a multilevel data structure. Within the traditional multilevel modeling framework, student scores are the level-1 unit and are nested within teachers, which is level 2, and finally nested within schools, which is level 3. These groupings of schools are often referred to as "clusters"; however, the appropriate level of aggregation varies depending upon the research questions and hypotheses.

Before a researcher moves forward with a hierarchical data analysis, the researcher must determine the amount of dependency within the groups by calculating the intraclass correlation coefficient (ICC). The ICC measures the proportion of variance in the outcome variable that can be explained at the between-group or cluster-level (Raudenbush & Bryk, 2002). The ICC for a two-level model, students nested in teachers, is expressed as

$$\rho = \frac{\tau_{00}}{(\sigma^2 + \tau_{00})}. \quad (1)$$

In Equation 1, ρ represents ICC, τ_{00} represents between group (level-2) variance and σ^2 represents within-group (level-1) variance. A high ICC indicates there is a lot of dependency within groups, and the hierarchical structure of the data should be accounted for during data analysis. If the ICC is 0 it indicates that none of the variability in the outcome variable is due to between-group differences and the hierarchical structure of the

data can be ignored. Even a low ICC indicates a dependency within groups and hierarchical linear modeling should be used.

Types of DIF

There are two types of DIF: uniform and nonuniform DIF. Uniform DIF occurs when the item in under consideration provides a constant relative advantage for the same group regardless of the trait level (Penfield & Camilli, 2007). In nonuniform DIF, one group may, for instance, have a relative advantage at low trait levels but a relative disadvantage at high levels (Penfield & Camilli). DIF can be examined by looking at the differences between focal and reference group item parameters, including item difficulty (location) for instance and item discrimination. The focal group is typically the group of interest, and the reference group is the standard or comparison group (Atar, 2007). However, the focal group can be thought of as the manifest group, which has the lower probability of obtaining the correct answer to or endorsing an item, and the reference group has a higher probability of getting the correct answer or endorsing the item (Wen, 2014). Finally, according to Quesen (2016), focal groups often have smaller sample sizes, whereas reference group populations are usually larger.

Traditional DIF Detection Procedures for Single-Level Data

Researchers often organize DIF detection procedures into model-based and non-model-based approaches. The most well-known non-model-based methods are the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993). Other well-known model-based DIF detection procedures are logistic regression (LR; Swaminathan & Rogers, 1990), Lord's χ^2 test (Lord, 1980), Raju's area approach (Raju, 1988, 1990), and the likelihood ratio

test (Tay, Meade, & Cao, 2015). Two commonly used DIF detection procedures are the MH procedure and LR procedure with ordinal data.

Mantel-Haenszel (MH) procedure. The MH procedure is one of the most widely studied and used methods for studying DIF in dichotomous items. It is a non-model-based contingency table method for detecting test items that performs differently between groups of test takers. The MH procedure and its extensions are relatively easy to calculate, do not require large sample sizes (e.g., 200 examinees per group) when working with non-nested data structures (Clauser & Mazor, 1998), and have an associated test of significance (Wood, 2011). The MH procedure is highly efficient in terms of statistical power and computational requirements (Clauser & Mazor). When using the MH procedure, researchers should assume that test takers are comparable, meaning they know the same amount of information; therefore, they will perform in much the same way on a specific item, regardless of their group membership (Holland & Thayer, 1986).

Ordinal logistic regression (OLR) procedure. Another popular DIF detection procedure is the LR procedure. As early as the 1990s, LR was being suggested as a DIF detection procedure. Miller and Spray (1993) described the use of LR, or the cumulative logit model (Agresti, 2007), for DIF detection. The separate, cumulative logits, Miller and Spray stated, can be incorporated into one model called a proportional odds model (Agresti). The LR procedure can identify uniform and nonuniform DIF, whereas the generalized Mantel-Haenszel procedure is only suited for identifying uniform DIF.

When item responses are ordinal, ordinal logistic regression (OLR) can also be used for detecting DIF. The OLR model is as follows (Scott et al., 2009, Equation 2):

$$ln[\frac{Pr(Y \leq k|g,\theta)}{1-Pr(Y \leq k|g,\theta)}] = \beta_{0k} + \beta_1\theta + \beta_2 g + \beta_3(g\theta). \quad (2)$$

In Equation 2, Pr (Y ≤ k) represents the probability of responding to an item in category k or below (for k = 0, 1, 2), θ represents ability, which is measured by the total score, g is the grouping variable (0 for reference, 1 for focal group), $g\theta$ is the interaction term between the grouping variable and ability, and β_{0k}, β_1, β_2, and β_3 are constants. When testing for uniform DIF, only two models need to be compared. The baseline model (R_1), which only includes the ability term, whereas the larger model (R_2) includes the ability term plus the grouping variable as predictors. The value of the difference in −2 log-likelihood of full R_2 and R_1 is used to detect DIF and tested for significance by using a Chi-square distribution with one degrees of freedom. If this comparison yields a significant result, the item is flagged for uniform DIF.

Multilevel DIF Detection Procedures

Multilevel regression models focus on the variability across levels of nested data. Research suggests that omitting levels of nesting during analysis leads to biased standard error estimates, inaccurate parameter estimates, inflated effect size estimates, and an increased risk of committing a Type I error (Snijders & Bosker, 2012). When data are sampled hierarchically, the observations are often not independent, which is a common assumption in many statistical analyses. For instance, if a researcher takes a sample consisting of students nested within schools and omits the school-level structure, the researcher assumes there is no similarity among students within those schools, and the

schools have no explanatory impact on the outcome variable of interest. Flawed tests of significance may occur, and researchers may conclude that there are true effects (impacts) present when only a sampler error differences exists (Nolan, 2016).

Multilevel generalized Mantel-Haenszel (MGMH). In 2013, French and Finch stated that few evaluations of a multilevel version of MH for DIF detection procedure to account for nested data had been conducted. In the years that followed, French, Finch, and Imekus (2019) tested the procedure a second time. The researchers proposed a method to account for multilevel data based on an adjusted test statistic that accounts for higher level covariance (Begg, 1999). Begg (1999) modified the MH because it only works with binary outcomes and does not follow a chi-square distribution if the observations are correlated. The modified Begg MH (BMH) method involves estimating the variance in the MH statistic caused by clustering, in addition to the ordinary variance that assumes no clustering (Begg, 1999). The BMH adjusts the MH statistic using a factor based on the ratio of the score statistic variance estimated using logistic regression, which accounts for multilevel data using the generalized estimating equation (GEE), to the ordinary variance of the score statistic, which does not account for multilevel data (French & Finch, 2013). The ordinary and GEE-based logistic regression models are both expressed as (French & Finch, 2013)

$$\log\left(\frac{P_{ki}}{1-P_{ki}}\right) = \beta_0 + \beta_1 X_i + \beta_2 Y_{i.} \quad (3)$$

In Equation 3, P_{ki} is the probability of a correct response to item k, β_0 is the intercept, X_i is the group membership for subject i, Y_i is the matching subtest score for subject i, β_1 is the coefficient for group variable, and β_2 is the coefficient for matching subtest variable.

GEE models are useful when finding population average effects of a covariate and not the individual specific effect, whereas multilevel modeling allows researchers to find the estimates of the varying coefficients, particularly varying slopes. Marginal models, such as GEE, answer different questions than conditional models, such as traditional multilevel models. For example, a marginal or population-average model might answer a question regarding the probability of an event in general; however, the conditional model, subject-specific models, would answer a question regarding the probability of an event for people in different situations.

The score statistic using this method tests the null hypothesis of no association between the predictor variable(s) and the response. The ordinary (not accounting for clustering) and GEE (accounting for clustering) models differ in how the covariance of the response is handled with respect to clustering. In the ordinary approach, the covariance matrix for the response with respect to clusters is the identity matrix, in which the off-diagonal elements are 0, which is equivalent to stating that the ICC is equal to 0. However, the GEE model does not assume an identify covariance matrix, but estimates the off-diagonal elements, which is an unstructured covariance matrix. In an unstructured covariance matrix, a unique covariance is estimated for each cluster. For each model, the variance of the score statistic is obtained using this covariance matrix. The ratio of these variances is expressed as

$$f = \frac{\sigma^2_{GEE}}{\sigma^2_{Ordinary}}. \quad (4)$$

In Equation 4, $\sigma^2_{GEE}$is the GEE-adjusted variance of the score statistic, accounting for clustering, and $\sigma^2_{Ordinary}$ is the ordinary variance of the score statistic, ignoring clustering.

The adjusted MH statistic (MH_B) can be used to analyze data gathered under a cluster sampling design (Begg, 1999) and is expressed as (French & Finch)

$$MH_B = \frac{MH}{f}. \quad (5)$$

In Equation 5, *MH* is the standard Mantel–Haenszel χ^2 test statistic. When there is no correlation in scores among test takers from the same cluster, such as a school, $f = 1$ and $MH_B = MH$. When the within-cluster correlations are large, σ^2_{GEE} will be larger than $\sigma^2_{Ordinary}$, leading to an *f* value that is relatively large and positive. Thus, this large positive *f* value serves to reduce the size of MH_B, indicating that the ordinary *MH* is an overestimate because it ignores clustering. Thus, MH_B is penalized for the degree to which clustering or dependency matters. Of note, the MH_B is not a true multilevel statistic, instead the MH_B is the *MH* adjusted for the degree to which the data deviate from ICC = 0.

Hierarchical logistic regression (HLR) and hierarchical ordinal logistic regression (HOLR) with multilevel data. To account for the nested structure in a dataset, HLR and HOLR have been used as DIF detection procedures with both dichotomous and polytomous data, respectively. HOLR models are used with multilevel data to predict an ordinal dependent variable measured on a Likert-type scale based on one or more independent variables. Consequently, the single-level DIF detection procedure can be extended to accommodate multilevel data.

HOLR models can be used with multilevel data to predict an ordinal dependent variable measured on a Likert-type scale based on one or more independent variables. The single-level DIF detection method can be extended to accommodate multilevel data. The general model for the logit of responding at or below category *k* to an item for the *i*th person (e.g., student) in the *j*th cluster (e.g., school) for two levels can be expressed as (Sharafi et al., 2017; Equation 2):

$$
\begin{aligned}
&level\ 1\text{: } \eta_{ij} = \ln\left[\frac{p\left(Y_{ij} \le k \middle| X_{qij}, W_{sj}\right)}{1-p\left(Y_{ij} \le k \middle| X_{qij}, W_{sj}\right)}\right] \\
&\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{qj}X_{qij} \qquad (6) \\
&level\ 2\text{: } \beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{qs} + u_{qj}
\end{aligned}
$$

In Equation 6, Y_{ij} is the polytomous item response for person *i* in cluster *j*. The *X*s represent level 1 level predictors, whereas *W*s are cluster level predictor. β and γ are the associated regression coefficients for *X* and W, respectively, and u_{qj} is the random effects at level 2. This general model for uniform DIF for within cluster variables can be expressed as (Sharafi et al., 2017; Equation 3):

$$
\begin{aligned}
&level\ 1\text{: } \eta_{ij} = \ln\left[\frac{p\left(Y_{ij} \le k \middle| \theta_{ij}, G_{ij}\right)}{1-p\left(Y_{ij} \le k \middle| \theta_{ij}, G_{ij}\right)}\right] \\
&= \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}G_{ij} \qquad (7) \\
&level\ 2\text{: } \beta_{0j} = \gamma_{00} + \mu_{0j}
\end{aligned}
$$

In Equation 7, Y_{ij} and θ_{ij} are the polytomous item response and ability for person *i* in cluster *j*. G_{ij} is the group identifier. If the model in Equation 7 is significant when

compared with the baseline model which does not include the group identifier, then the studied item will be flagged as showing uniform DIF.

Multilevel Applied and Simulation DIF Detection Studies

A limited number of applied or simulation studies have considered multilevel DIF detection procedures, for either dichotomous or polytomously scored data. Those identified works are described below.

Substantive studies examining three-level models for DIF detection using dichotomous items. Eight applied studies examined three-level models for DIF detection using dichotomous items between 2005 and 2018. Kamata, Chaimongkol, Genc, and Bilir (2005) used a random-effects model to estimate DIF across groups. They analyzed data from the 2003 NAEP 4th grade mathematics assessment and examined DIF between the limited English proficiency sample who received test accommodations and those students who did not. DIF between accommodated and non-accommodated students was detected, and the variations of the magnitude of DIF across schools was estimated.

Cheong (2006) studied the effects of school context on DIF in a large-scale assessment. He used an HGLM framework to detect DIF and identify school-level variables that might cause DIF. He illustrated the method using civic items to determine if they contained ethnic–racial DIF. Cheong's study had a three-level model: items within persons within schools.

Although Kamata et al. (2005) and Park (2008) examined DIF in mathematics, they used different sources of data and different procedures. Park investigated a modeling approach using multilevel IRT for cross-national comparisons. He illustrated the application in a study of the Trends in International Mathematics and Science Study

(TIMSS) 2003 grade 8 mathematics assessment. He used a 3-level model: item-level, student-level, and country-level.

Like Park, Burkes (2009) used TIMSS data to investigate DIF. She identified socioeconomic status (SES) differences in student performances on the 2003 TIMSS eighth-grade mathematics assessment, using Kamata and Binici's (2003) multilevel-DIF methodology. She identified mathematics items that functioned differently in high and low SES students with similar ability. Burkes had a three-level model: item-level, student-level, and classroom-level.

Beaver, French, Finch, and Ullrich-French (2014) used a multilevel MH DIF procedure to examine sex differences in dichotomous item responses for examiner ratings of children's social-emotional skills on the Brigance Inventory of Early Development III SE scale. Children were nested within childcare sites. They found that scores did not appear to be influenced by rating distortions based on sex stereotypes.

Finch, Finch, and French (2016) conducted a different cross-national study. They investigated DIF in the Progress in International Reading Literacy Study (PIRLS) items across multiple European countries. They used a multilevel LR-type technique with students nested within nations (clusters). They were interested in the extent to which a mother's primary language was associated with DIF on reading items and whether these relationships were consistent across countries. They showed that DIF based on the mother's language was present for several items, but patterns of DIF differed across nations.

French, Finch, and Vazques' 2016 study is similar to the one conducted by Beaver et al. (2014) in that they both used Brigance data. However, French et al. (2016)

investigated a multilevel version of SIBTEST (Shealy & Stout, 1993) to illustrate DIF detection in a multilevel context with dichotomous items. The authors used national data from the Brigance Comprehensive Inventory of Basic Skills – II mathematics assessment between boys and girls, and students were clustered within schools. They found that adjusting DIF statistics for clustered data resulted in fewer items flagged for DIF compared to no adjustments.

French, Finch, Randel, Hand, and Gotch (2016) chose to study DIF in critical thinking rather than a traditional academic content area. They outlined a method for evaluating measurement sensitivity by conducting content and DIF analyses to detect intervention effects and test for measurement sensitivity. They collected data in a multilevel framework with students nested in classrooms from the Cornell critical thinking tests, which was scored dichotomously. They used the multilevel MH as a DIF detection procedure. Their results suggested that although mean differences were not observed across all content domains, there were intervention effects associated with some assessment items.

The French et al. (2016) study was implemented in two distinct steps, the first engaging experts with a teacher-focused professional development intervention to conduct a content analysis to align items on a general assessment with the intervention. The second step used DIF analysis to test the sensitivity of items identified in step 1 that were related to the intervention. Data were taken from a randomized cluster field trial that used the science writing heuristic (SWH) approach to learning science. The SWH is an immersive approach to teaching scientific argument and is examined in a randomized control trail study in the Midwest with comparison and treatment schools. Participants

were a representative sample of 2,181 treatment and 1,004 control students taken from 48 schools in a Midwestern location in grades 3 through 5.

These eight applied studies conducted analyses in various academic contexts, including mathematics, civics, critical thinking, science writing, and reading, looking for items that exhibited DIF based on gender, disability, teaching method, and language. Their models used a basic format of item, within person within school.

Simulation studies examining three-level models for DIF detection using dichotomous items. Eight simulation studies examined 3-level DIF detection using dichotomous items between 2010 and 2015. These studies appear to be a disparate collection of random studies, but they are anything but – they are united by performance questions based on Type I error and power and manipulated simulation conditions

An example of a simulation study with dichotomous items is the one conducted by French and Finch (2010), in which they evaluated two DIF techniques, standard LR and HLR that accounts for multilevel data. They simulated data using a hierarchical framework, such as examinees clustered in schools. The authors found that when the grouping variable was within clusters, LR and HLR performed equally well in terms of Type I error control and power. However, when the grouping variable was between clusters, standard LR failed to maintain the Type I error rate of .05.

Another example of a simulation performed with dichotomous items is illustrated by Patarapichayatham, Kamata, and Kanjanawasee (2012). This team examined cross-level, two-way DIF models for dichotomously scored items in a Rasch item response model. Their simulation study demonstrated that the quality of parameter estimates can be affected by model selection strategies and certain simulation conditions. They found

that when cluster-level DIF and cluster size became larger, all model selection strategies tended to select the most complete model, However, when the effects of cluster size were smaller, this was not necessarily true.

Next, French and Finch (2013) investigated the effectiveness of several DIF detection procedures with nested data (examinees nested in schools) using the multilevel MH procedure. They used the 2PL model to simulate data. They found that the multilevel MH procedure was preferable to the standard MH in the presence of multilevel data, mainly when the ICC was relatively large, over .25.

Jin, Meyers, and Ahn (2014) compared the performance of DIF detection procedures when the ICC of the studied item (ρ_y) was less than the ICC of the total score (ρ_x), which is commonly found in practice. The performance of two DIF detection procedures that do not account for multilevel data structure, MH and LR, was compared with HLR when $\rho_y < \rho_x$, which have not been studied. They found that when the grouping variable was at the cluster level, HLR, LR, and MH performed equivalently in terms of controlling Type I error rate at the nominal alpha level 0f .05 when ρ was small (i.e., .25) under both item generating models, Rasch and 2PL. When ρ became larger (i.e., ρ = .25), HLR generally outperformed LR regarding Type I error rate, and MH was slightly conservative under both models. HLR, LR, and MH maintained power above the acceptable level (.80) with trivial differences across all levels of manipulated factors.

Wen (2014) conducted DIF analyses with multilevel data using a simulation that emphasized DIF at the cluster-level only and DIF at the student and teacher levels. Wen extended Kamata's (2001) three-level Rasch model by adding covariates, allowing him to understand the factors that affect DIF detection or the impact of DIF on ability estimation

more completely. Wen's simulation showed that the estimates of fixed parameters were close to true values, indicating the multilevel Rasch model is reliable in terms of DIF detection.

Like Wen, Francis (2015) based his study on Kamata's previous work. Francis investigated the performance of two models for nested item response data, using Kamata's multilevel IRT. He examined the causes of DIF, specifically if DIF was present at the cluster-level. He used a four-level longitudinal logistic regression model with the nesting pattern of items, time points, students, and schools. His simulation showed that the model for DIF detection was powerful and accurate in identifying DIF at the item and school levels and that sample size had a significant effect on DIF detection at both levels.

Unlike other researchers, French and Finch (2015) investigated uniform and nonuniform DIF. They examined the performance of multilevel adaptations of SIBTEST with multilevel data to examine Type I error and power rates. Dichotomous data were generated using the 2PL item response model, and students were nested within clusters. Results showed that for both uniform and nonuniform DIF detection, ignoring the multilevel data structure will likely yield inflated Type I errors, which could lead to an incorrect determination of DIF.

In 2018, Shear conducted a simulation to demonstrate and evaluate a random coefficient hierarchical logistic regression model to test for uniform DIF and DIF variance. Item responses were generated via a 2PL model, and examinees were clustered in groups. He found that the model is a promising approach to understanding DIF.

Through the use of simulation, these eight groups of researchers employed and/or compared the performance of DIF detection techniques based on Type I error and

statistical power. They used multiple types of data generating models, such as a three-level Rasch and 2PL. Their DIF detection procedures included LR, HLR, MH, and the SIBTEST. The researchers also varied where DIF was occurring, within or between clusters. Their results helped clarify the appropriate use of DIF detection procedures with multilevel data.

Substantive studies examining three-level models for DIF detection using polytomous items. Only one applied study was found that examined DIF in a 3-level model with polytomous items. Finch and French (2010) tested for uniform DIF between male and female students on end-of-semester class evaluations in a university science course and demonstrated DIF detection procedures that accounted for nested data. They used the multiple indicator multiple cause (MIMIC) model, which allows for the representation of a latent variable using multiple indicators, such as items on a survey. This research demonstrated the flexibility of analyzing such data using the MIMIC and the hierarchical MIMIC model, which allows for the inclusion of individual and group-level variables.

Simulation studies examining three-level models for DIF detection using polytomous items. Three simulation studies examined three-level models for DIF using polytomous items between 2006 and 2017. Vaughn (2006) investigated DIF for polytomous items from a 3-level (item, person, and cluster) LR perspective using the GRM as a generating model. His first simulation used a fixed multilevel DIF model, whereas the second simulation applied a random DIF model. Results showed that all parameter estimations in the fixed and random DIF models had little bias.

Sharafi et al.'s 2017 had features in common with Vaughn's. They evaluated the effectiveness of two DIF detection procedures in nested polytomously scored data generated by a multilevel GRM. The authors used OLR, which only accommodates level 1 information, and HOLR to assess DIF in simulated and empirical multilevel polytomous data. They found that HOLR and OLR performed almost equally in terms of controlling Type I error rate at the alpha level of .05.

Unlike Vaughn (2006) and Sharafi el al. (2017), French et al. (2019) did not use LR-related procedures to examine DIF but instead used MH-related techniques. French et al. investigated the performance of the GMH procedure and a MGMH procedure for the detection of uniform DIF with multilevel data and polytomous items. Multilevel data were generated with manipulated factors, including intraclass correction, subjects per cluster, to examine Type I error rates and power. Their results showed the differences in DIF detection when the analytic strategy matches the data structure. Specifically, the GMH had an inflated Type I error rate across conditions, and therefore, artificially high power. On the other hand, the MGMH had good power rates and maintained the Type I error rate.

These three studies are tied together by their use of polytomous items to investigate DIF but differ in the techniques the researchers employed. While the first sets of researchers used LR-type procedures, French et al. used Mantel-Haenszel-related techniques to explore DIF.

Purpose

French et al. (2019) recognized that DIF detection procedures, such as MH and LR, have been evaluated for dichotomously scored test items; however, they also

acknowledge there is a gap in the research literature on the use of the MGMH procedure for DIF detection with polytomously scored items in a multilevel framework. Their study examined the performance of the Begg (1999) adjusted methods for MH with polytomous items, building on work done with dichotomous items. French et al. (2019) simulated data for their study and used the GRM to generate data.

Although one might find a compendium of research papers on DIF simulation studies with polytomous items to be thin at best, to find a 3-level (item, participant, school) DIF simulation study with polytomous items generated by the PCM would indeed be a rarity. At a time when large-scale assessment development is in transition, the assessment development community must have a comprehensive repertoire of reliable analysis techniques, for DIF detection and other procedures, available for both dichotomous and polytomously scored items at both the single-level and multilevel.

The purpose of this Monte Carlo simulation is to evaluate the Type I error rate and Power of two multilevel DIF detection procedures, HOLR and the MGMH, with data simulated under various conditions in a multilevel data structure perspective. While the generalized MH (GMH) is one of the most proven methods of DIF detection for polytomous item response data, support for the MGMH is just beginning to accumulate (French et al., 2019, French & Finch, 2013). In addition, although these methods have been examined under some simulation conditions, they have never been directly compared. Findings from this study will directly benefit practitioners who work with hierarchical polytomously scored data. The fundamental research question is one of DIF detection procedure performance efficacy: How do the two DIF detection procedures, MGMH and HOLR, compare in terms of performance efficacy measured by Type I error

and statistical power? Using a multilevel version of the PCM (Master, 1982) to generate data, the accuracy of the MGMH and HOLR DIF detection procedures with respect to Type I error and power will be evaluated.

The significance of this study lies in the use of the MLPCM to generate data, the further investigation of MGMH and its comparison to HOLR, and addition to the knowledge base of HOLR as a DIF detection procedure.

Chapter Two: Method

Simulation Design

Data were generated using a multilevel version of Master's partial credit model (1982). Four factors were manipulated: number of participants per cluster (20, 40; Chaimongkol, 2005; French & Finch, 2010, 2013; Jin et al., 2014), number of clusters (50, 100, 200; French & Finch, 2010), DIF magnitude (0, .4, .8; French & Finch, 2010; Garrett, 2009; Sharafi et al., 2017; Su & Wang, 2005; Wen, 2014), and ICC (.05, .25, .45; French & Finch, 2010, 2013; Sharafi et al., 2017) resulting in 72 conditions. Test length (20 polytomous items; Dodeen, 2004; Dodeen & Johanson, 2001; French & Finch, 2010; Finch & French, 2007; Garrett, 2009; Sharafi et al., 2017), number of response categories (5), number of items with DIF (10% or 2; Chang, Mazzeo, & Roussos, 1996; Su & Wang, 2005; Wang & Su, 2004; Zwick, Thayer, & Mazzeo, 1997), type of DIF (uniform; Williams, 2003; Zwick, Donoghue, & Grima, (1993), data generating model (MLPCM), grouping variable (level 1 within cluster, French & Finch), and balanced sample size ratios (SSR) between focal and reference groups (10:10, 20:20; French & Finch, 2010) were kept constant. The number of replications per condition is 400. Table 1 shows the simulation variables.

Table 1

Simulation Variables

Variable	Value
Manipulations	
Number of participants per cluster	10, 20, 40
Number of clusters	25, 50, 100
ICC	.05, .25
Magnitude of DIF	0, .2, .4, .8
Constant	
Number of items	20
Type of items	Polytomous with five response categories
Data generating model	MLPCM
Type of DIF	Uniform
Location of DIF	Within cluster – level 1
Grouping variable - dichotomous	Level 1: Within cluster
Proportion of DIF	10% or 2 items
Theta distribution	Normal (0, 1)
Sample size ratio	Balanced between reference and focal groups: 10:10 and 20:20

Note. DIF = differential item functioning; MLPCM = multilevel partial credit model; ICC = intraclass correlation coefficient.

Data Generation

Item parameters. Table 2 shows the item parameters that were used in the MLPCM, which are based on item parameters used by Wang and Shih (2010) and Wang and Su (2004). For the rest of the items in the study, the parameters will be duplicated.

Table 2

Item Parameters for Data Generation

Item	α_i	δ_i	τ_{i1}	τ_{i2}	τ_{i3}	τ_{i4}
1	1	0.81	-1.16	-0.29	0.32	1.13
2	1	1.07	-0.89	-0.33	0.35	0.87
3	1	0.72	-1.09	-0.69	0.20	1.58
4	1	0.58	-1.14	-0.71	0.22	1.64
5	1	0.87	-1.25	-0.38	0.17	1.46
6	1	0.93	-1.54	-0.30	0.44	1.41
7	1	1.05	-1.04	-0.38	0.28	1.13
8	1	0.88	-1.11	-0.57	0.10	1.58
9	1	1.00	-1.31	-0.40	0.27	1.44
10	1	0.93	-1.29	-0.40	0.27	1.41

Note. α_i = slope (discrimination) of item i, δ_i = difficulty of item *i*, $\tau_{i1} - \tau_{i4}$ = category thresholds for item *i*. Reprinted from Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166-180.

Software

Data generation was conducted in R (R Core Team, 2019). The MGMH data analyses were conducted in SAS 9.4 (SAS Institute, 2013), and the HOLR analyses were conducted in R (R Core Team, 2019). See Appendix A for the R code that generated the

data. Appendix B contains the code for the MGMH analysis (French, Finch & Iverson, 2015) and Appendix C contains the code for the HOLR analysis.

Evaluation Criteria

Type I error rates and power. The dependent variables in this study are the Type I error rate and statistical power of the two DIF detection procedures, HOLR and MGMH. Items generated to have DIF were examined for power, and Type I error rates were examined for non-DIF items. Type I error rate is the percentage of time a DIF detection procedure flags an item for DIF when it does not contain DIF. Statistical power is the percentage of time the DIF detection procedure flags an item for DIF when the studied item contains DIF or the proportion of cases in which DIF items are correctly identified. Values that are equal to or larger than 0.8 are thought to indicate high power. Type I error rates were calculated by assuming that the parameters of the studied item were identical for the focal and reference groups. Both Type I error rates and power were calculated within each condition and across all replications. Type I error rates were evaluated at the .05 level.

Chapter Three: Results

Non-convergence

During data analysis for the MGMH DIF detection procedure, a non-convergence issue occurred. It is likely that the MGMH and Begg adjustment conducted relied on the inversion of a matrix that could not be performed with some random datasets and was more likely in the conditions specified.

The data were organized in the following manner. A total of 72 folders were created, representing 72 simulated conditions. An abbreviated outline of the first six folders is shown in Table 3 to illustrate the format.

Table 3

Organizational Structure of Data for First Six Folders

Folder	Number of clusters	Persons per cluster	DIF	ICC
1	25	10	0	.05
2	50	10	0	.05
3	100	10	0	.05
4	25	20	0	.05
5	50	20	0	.05
6	100	20	0	.05

Note. DIF = differential item functioning; ICC = intraclass correlation coefficient.

In each of the 72 folders, 400 files were stored that represented the 400 replications for each condition. This means that 28,800 data files (72 x 400) were created. The creation of these files was not a smooth, seamless process. At times, four computers were engaged to create the necessary files. Often, random test sets were not analyzed for no apparent reason and had to be re-analyzed.

When all 72 folders contained 400 files, meaning all 28,800 data files were created, a final check was made. Each of the 400 files contains 20 items with associated p-values for a total of 576,000 *p*-values (72 x 400 x 20). At this time, it was found that a seemingly random selection of the 400 files were completely empty, while others were only partially empty, with only a few missing p-values. Upon further inspection, one hundred fifty-five (155) files were found to be completely empty. Therefore, those

associated 155 test sets were re-analyzed creating the appropriate 3,100 *p*-values (155 x 20).

Even though the vast majority of files had a full complement of p-values, unfortunately, random files still with missing p-values remained, as shown in Table 4. The missing *p*-values occurred in conditions with 25 clusters, 10 persons per cluster at all levels of DIF, and both levels of ICC. In the final compilation for the MGMH analysis, 290 *p*-values were missing or .0053% (290/576,000). Of the 72 conditions, 8 (11.1%) had missing *p*-values (i.e., 25 clusters, 10 persons per cluster, all levels of DIF, both levels of ICC). Although these non-convergence issues are small, they could affect the results for Type I error and power associated with these 8 conditions.

Table 4

Folders and Files with Used and Missing p-values

Folder	Number of clusters	Persons per cluster	DIF	ICC	Files with missing p-values	Number of p-values used (missing)
1 (1-400)	25	10	0	.05	25	375 (25)
10 (3,601-4,000)	25	10	.2	.05	33	367 (38)
19 (7,201-7,600)	25	10	.4	.05	34	366 (34)
28 (10,801-11,200)	25	10	.8	.05	66	336 (66)
37 (14,401-14,800)	25	10	0	.25	35	366 (35)
46 (18,001-18,400)	25	10	.2	.25	19	381 (19)
55 (21,601-22,000)	25	10	.4	.25	36	364 (36)
64 (25,201-25,600)	25	10	.8	.25	34	366 (37)

Note. DIF = differential item functioning; ICC = intraclass correlation coefficient

Of the 18 conditions used to determine Type I error rates, the first 2 conditions (i.e., 25 clusters, 10 persons per cluster, 0 DIF, .05, .25 ICC) could have been affected by this non-convergence issue even though the number of missing p-values is quite small.

Type I Error Rate

The Type I error rates for the two DIF detection procedures, HOLR and MGMH, for all simulated test sets are summarized in Table 5. As can be seen from the table, across all 18 conditions, the Type I error rate for the HOLR DIF detection procedure was closely maintained to the nominal .05 level (ranged from .053 to .059). Sample size (number of clusters times number of participants per cluster) had little to no effect on Type I error for the HOLR DIF detection method. The mean Type I error for 250 participants was .055, .057 for 500, .055 for 1000, .055 for 2000, and .055 for 4000 participants.

The ICC has a small but perceptible impact on the Type I error rate for the HOLR DIF detection procedure. For example, within a pair of Type 1 error rates, the one with the higher ICC, has the higher Type I error rate. Note the comparison between two samples of 25 clusters with 10 persons per cluster but one with an ICC of .50 and the other with an ICC of .25. The first sample has a Type I error rate of .0534 and the second has a slightly higher error rate of .0578. This pattern of higher Type I error rates with larger ICCs is repeated throughout the data for the HOLR DIF detection procedure.

Table 5 also shows Type I error rates for the MGMH DIF detection procedure. Across all 18 conditions, Type I error rates for the MGMH procedure ranged from .010 to .130. Type I error rates tended to be inflated when the ICC was .05. Type I error rates for

the MGHM were at or above the nominal level of .05 for 13 of 18 (72%) conditions, with lowest rates occurring at ICCs of .25.

The mean Type I error rate for the HOLR procedure is .0594, whereas the mean value for the MGMH is .0754. Maximum difference in Type I error rates between the two procedures is |.0754| at 25 clusters times 40 persons per cluster and an ICC of .05. Minimum difference is |.0048| at 25 clusters times 20 persons per cluster and ICC at .25. Finally, the average difference between the two rates is -0.0205.

Table 5

Type I Error Rate for Two DIF Detection Procedures by Number of Clusters, Number of Persons per Cluster, Magnitude of DIF, and ICC

Number of clusters	Persons per cluster	DIF	ICC	HOLR	MGMH
25	10	0	.05	.053	.121
25	10	0	.25	.058	.086
25	20	0	.05	.059	.125
25	20	0	.25	.055	.050
25	40	0	.05	.055	.130
25	40	0	.25	.054	.023
50	10	0	.05	.055	.104
50	10	0	.25	.057	.066
50	20	0	.05	.053	.108
50	20	0	.25	.057	.033
50	40	0	.05	.055	.106
50	40	0	.25	.052	.011

100	10	0	.05	.055	.100
100	10	0	.25	.054	.061
100	20	0	.05	.057	.102
100	20	0	.25	.052	.028
100	40	0	.05	.052	.095
100	40	0	.25	.058	.010

Note. HOLR = hierarchical ordinal logistic regression; MGMH = multilevel generalized Mantel-Haenszel.

Power

Power for the two DIF detection procedures, HOLR and MGMH, for all simulated test sets are provided in Table 6. As can be seen from the table, across all 54 conditions for items 19 and 20, power for HOLR ranged from .250 to 1. Power increases as DIF magnitude increases, reaching 1 for a DIF magnitude of 0.80 and as sample size increases (number of clusters times persons per cluster). In many conditions, power was above the acceptable rate of .80. This included the following conditions:

- Lower sample sizes (N = 250, 500), DIF at .80, ICC at .05, .25
- Medium sized samples (N = 1,000), DIF at .40, .8; ICC at .05, .25
- Large sized samples (N = 2,000, 4000) DIF at .20, .40, .80; ICC at .05, .25

Not all conditions elicited high power, especially those with small sample sizes (i.e., 250, 500). For example, under the conditions of

- Sample size 250, DIF magnitude .20, ICC .05 power = .250,
- Sample size 250, DIF magnitude .20, ICC .25 power = .210,
- Sample size DIF magnitude of 0.4 and ICC .05 and .25, power = 0.690, .6750,

- Sample size = 250, DIF magnitude = 0.4, ICC = .05, power = .690
- Sample size = 250, DIF magnitude = 0.4, ICC = .25, power = .690
- Sample size = 500, DIF magnitude = 0.4, ICC = .05, power = .935
- Sample size = 500, DIF magnitude = 0.4, ICC = .25, power = .910
- At sample sizes of 2,000 and 4,000, power is 1.

Table 5 also shows power for the MGMH DIF detection procedure. Across all 54 conditions, power for the MGMH DIF detection procedure ranged from .1350 to 1. Statistical power increases for the MGMH procedure as sample size (number of clusters by persons per cluster) increases. Power was found to be consistently high with sample sizes over 1,000.

DIF magnitude appears to have an impact on power but that impact is not consistent, especially with smaller sample sizes. Moreover, the interaction of DIF magnitude and ICC influences power for the MGMH. There does seems to be a consistent trend of lower power with higher ICCs that is relatively consistent throughout the data. This is true for both detection procedures, but it is not present at the largest sample sizes. The following examples are given to illustrate this point.

- At a sample size of 500 or 1,000, power is greater than .80 with a DIF magnitude of .40 if the ICC is .05 but not if ICC is .25.
- At smaller sample sizes (e.g., N = 250), DIF magnitude = .80, and ICC = .05, power is .60. However, power is only .22 when ICC increased to .25.
- At sample size of 500 and DIF magnitude = .80, when ICC was .05 power was .8950 but was only .51 when ICC increased to .25.

The HOLR showed more values closer or greater to .80 compared to MGMH. Mean power, across all conditions, for the HOLR procedure is .8690, whereas the mean value for the MGMH is .7690. Maximum difference in power between the two procedures is 1.000 at 100 clusters times 20 persons per cluster, DIF = .8 and an ICC of .25. Minimum difference is .000 under most conditions. Finally, the average difference in power between the two detection procedures is .100.

Table 6

Power for Two Detection Procedures by Number of Clusters, Participants per Cluster, DIF Magnitude, and ICC for Items 19 and 20

Number of clusters	Participants per cluster	DIF	ICC	Power HOLR	Power MGMH
25	10	0.2	.05	.250	.235
25	10	0.2	.25	.210	.135
25	10	0.4	.05	.690	.380
25	10	0.4	.25	.675	.220
25	10	0.8	.05	1.000	.600
25	10	0.8	.25	1.000	.220
25	20	0.2	.05	.410	.000
25	20	0.2	.25	.400	.250
25	20	0.4	.05	.935	.600
25	20	0.4	.25	.910	.375
25	20	0.8	.05	1.000	.895
25	20	0.8	.25	1.000	.510
25	40	0.2	.05	.685	.585

25	40	0.2	.25	.655	.500
25	40	0.4	.05	1.000	.895
25	40	0.4	.25	1.000	.720
25	40	0.8	.05	1.000	1.000
25	40	0.8	.25	1.000	.970
50	10	0.2	.05	.400	.520
50	10	0.2	.25	.370	.380
50	10	0.4	.05	.935	.800
50	10	0.4	.25	.925	.635
50	10	0.8	.05	1.000	.990
50	10	0.8	.25	1.000	1.000
50	20	0.2	.05	.680	.845
50	20	0.2	.25	.685	.680
50	20	0.4	.05	1.000	.980
50	20	0.4	.25	.995	1.000
50	20	0.8	.05	1.000	1.000
50	20	0.8	.25	1.000	1.000
50	40	0.2	.05	.920	.995
50	40	0.2	.25	.930	.970
50	40	0.4	.05	1.000	1.000
50	40	0.4	.25	1.000	.895
50	40	0.8	.05	1.000	.990
50	40	0.8	.25	1.000	.885

100	10	0.2	.05	.730	.870
100	10	0.2	.25	.685	1.00
100	10	0.4	.05	.995	1.00
100	10	0.4	.25	0.995	1.00
100	10	0.8	.05	1.00	1.00
100	10	0.8	.25	1.00	1.00
100	20	0.2	.05	0.930	1.00
100	20	0.2	.25	0.930	1.00
100	20	0.4	.05	1.00	1.00
100	20	0.4	.25	1.00	1.00
100	20	0.8	.05	1.00	1.00
100	20	0.8	.25	1.00	.000
100	40	0.2	.05	1.00	1.00
100	40	0.2	.25	1.00	1.00
100	40	0.4	.05	1.00	1.00
100	40	0.4	.25	1.00	1.00
100	40	0.8	.05	1.00	1.00
100	40	0.8	.25	1.00	1.00

Note. HOLR = hierarchical ordinal logistic regression; MGMH = multilevel generalized Mantel-Haenszel.

Chapter Four: Discussion

Implications

The importance of DIF and its potential impact on the validity of assessments, especially in an environment that relies heavily on standardized testing, should not be

understated. Consequently, practitioners and test developers should endeavor to ensure that test items and test scores accurately reflect the traits of test takers being measured.

Rightfully so, researchers have focused on procedures for detecting DIF; however, they have not paid as much attention on the procedures to explore DIF in multilevel data structures. As the importance of assessments that are "valid, reliable, and comparable for all students and for each subgroup of students among participating schools and districts" (U.S. Department of Education, 2017, p. 5) continues, increases, or expands across contexts, the importance of DIF detection procedures for educational tests that polytomous in nature continues to increase.

DIF analysis methodologies are psychometric tools used to confirm assessment fairness and validity and employing more than one method to analyze DIF is beneficial in confirming DIF results. Given its prominence in the test development process, it is imperative that DIF detection procedures are accurate. The results of this study help clarify and enhance several issues for test developers and psychometric practitioners. For example, this study contributes to the budding literature begun by French and Finch (2013) and French et al. (2019) on the effectiveness of adjusted statistical methods for DIF detection in the presence of multilevel data. In addition, this study contributes to the literature that examines DIF detection procedures for polytomously scored items (e.g., French and Miller, 1996) within a multilevel framework.

As additional work is done to evaluate DIF detection procedures with multilevel data, psychometric tools are refined, improved, and enhanced, thus sharpening their ability to provide accurate item and test development guidance and decisions. Certainly,

other students and psychometricians will pick up where this study ends to create better tools.

Type I Error

The purpose of this study was to compare the performance of two DIF detection procedures, HOLR and MGMH, by comparing Type I error rates and power. Results of this study showed that the HOLR DIF detection procedure maintains the Type I error rate of .05 better than the MGMH DIF detection procedure. In addition, the HOLR Type I error rates showed a smaller range than the MGMH rates.

In this study, using the HOLR DIF detection procedure, higher Type I error rates were found with smaller sample sizes and ICC has a small but perceptible effect on Type I error, but the range of values was very small. These results indicate that HOLR controlled the nominal Type I error rate of .05 reasonably well. The Type I error rate for the MGMH DIF detection procedure was at or below the nominal level of .05 for six (33.33%) conditions, and mean Type I error rate was .0754, whereas the mean Type I error rate for the HOLR procedure was .0594.

The results of this study are consistent with those researchers who have studied DIF using LR techniques but not consistent with results from the few studies using MGMH. In a study similar to this one, Sharafi et al. (2017) found that HOLR controlled the nominal Type I error rate of .05, and French and Finch (2010) found that both standard LR and HOLR maintained the nominal Type I error rate of .05 across all manipulated conditions when the grouping variable was at the within-clusters level as is the case in this study.

Findings from this simulation study are not entirely consistent with the findings of French et al. (2019), who used the same MGMH and Begg's adjustment method as the one used in this study. The manipulated factors used in the French et al. study were similar to those in this study. For example, five ICCs were used: .05, .15, .25, .35, and .45. The number of clusters simulated was 50, 100, and 200, whereas the number of subjects per cluster was 5, 15, 25, and 50. Four DIF magnitudes were simulated: 0, .4, .6, and .8, and uniform DIF was specified. Although the constants in the French et al. study were also similar, differences should be noted. French et al. simulated 20 items, each with four response levels. Note, however, that the items in this study had five response options. A purified scale score was used in the French et al. study, with only one targeted item, whereas this current study did not use a purified scale score and had two targeted items. Moreover, data in their study were simulated using a multilevel graded response model, and in this study a multilevel partial credit model with different threshold parameters and discrimination values was used. Last, French et al. used 1000 replications.

French et al. (2017) found that for the within-cluster condition, the condition used in this study, all DIF detection procedures reported Type I error rates at or below the nominal level. In this study, the MGMH reported Type I error rates at or above the nominal level of .05 for 13 of 18 (72%) conditions, with lowest rates occurring at ICCs of .25. More interestingly, French et al. found that Type I error rates for all DIF detection procedures increased with associated increases of the ICC. In this study, when comparing a pair of conditions, the highest Type I error rates occurred when the ICC was .05 and the lower Type I error occurred when the ICC was .25; under no conditions did the Type I error rate meet the nominal .05 level when the ICC was .05 but sometimes did when the

ICC was .25. Consider the ICC levels in following pairs; Type I error rate is always lower for the condition with a higher ICC:

- Number of clusters 25, participants per cluster 10, DIF = 0, ICC = .05 and .25, Type I error = .1206 and .0861, respectively.
- Number of clusters 25, participants per cluster 20, DIF = 0, ICC = .05 and .25, Type I error = .1250 and .0500, respectively.
- Number of clusters 25, participants per cluster 40, DIF = 0, ICC = .05 and .25, Type I error = .1300 and .0228, respectively.

French et al. (2019) found that the Type I error rates decreased as the number of subjects per cluster increased. In this study, the trends in Type I error rates associated with the MGMH procedure are harder to characterize; however, the mean of the error rates for the top 9 conditions with smaller sample sizes is .090, whereas the mean for the bottom 9 conditions with larger sample sizes is .061.

Power

In this study, it was found that power is better for HOLR than MGMH DIF detection procedure. HOLR showed more values closer or greater to .80 compared to MGMH. In general, for either DIF detection procedure, for any DIF magnitude, 0, .2, .4 or .8, a large sample size is needed for adequate power.

The findings regarding HOLR and sample size are consistent with other studies that showed power rates increase as sample size increase. However, the increase in power does not appear to be tied only to cluster size but to overall sample size. The findings in Table 6 show that for sample sizes at or over 1,000, 83.33% (30/36) of the conditions have power rates at or above .80. Other studies, such as French and Finch (2010) found

that power increase with increasing cluster size. Last, Sharafi et al. found that HOLR maintained power above the acceptable level (.80) across most of the studied conditions. Power was lower than .80 for sample sizes less than 500 with low level of DIF magnitude across all levels of ICC. They found that power was high for larger samples.

The MGMH power results are more consistent with results from previous studies than the MGMH findings for Type I error. In this study, power increases for the MGMH procedure as sample size (number of clusters by persons per cluster) increases. Power was found to be consistently high with sample sizes over 1,000. French et al. (2019) found that only when the ICC was .05 did the MGMH procedure report power estimates above the desired .80 level. However, in this study, power estimates reached .80 with an ICC of .25 only when sample sizes reached 1,000.

The influence of DIF magnitude in this study was consistent with that found by French et al. They found that power rates were lowest for the lowest level of DIF condition (i.e., .40). In this study, at lower sample sizes and ICC .05, power was .380 at DIF magnitude .4 and .600 at DIF magnitude 0.8. French et al. noted that only when the DIF magnitude was .60 did the MGHM procedure report statistical power above 0.80. In addition, only when the number of clusters was 100 or 200 did the MGHM report an acceptable level of power for DIF detection.

The most likely reason for the discrepancy in the findings between the French et al. (2019) and this study is the way the data were simulated. The data in their study were simulated using a multilevel graded response model, and in this study a multilevel partial credit model with different threshold parameters and discrimination values was used. The use of a purified sample also could have affected the results. Last, because the non-

convergence issue systematically affected all conditions with 25 clusters, 10 persons per cluster at all levels of DIF, and both levels of ICC, they could affect the results for Type I error and power associated with these conditions.

Limitations and Future Research

Five important limitations are worth noting. First, purification was not performed on the simulated data. The Mantel-Haenszel procedure distinguishes between DIF and item impact by comparing the odds for success between groups after conditioning on ability; however, this conditioning requires a valid criterion. Most of the time, an appropriate external criterion is not available, therefore practitioners typically use the total test score as the matching criterion. Holland and Thayer (1988) recommended a way of improving the matching criterion by using a two-step form of the Mantel-Haenszel procedure in which items identified as showing DIF are removed from the matching criterion for subsequent analyses. Clauser, Mazor, and Hambleton (1993) reported that the results for the two-step procedure were equal or superior to the single-step procedure in identifying simulated DIF items across conditions investigated.

Clauser, Mazor, and Hambleton (1993) found that the most substantial improvement was noted when the purification procedure was applied to data simulated to have focal and reference groups of equal ability. If there is relatively little DIF in the test, the advantage afforded with the difference associated with using the two-step procedure will be minimal; however, when the test contains more DIF, the advantage will be greater. Since this simulation study only included two items, which is minimal (10% of the test), we chose to not use a purification approach.

Researchers have suggested the use of purification for use with LR DIF detection procedures (Rogers & Swaminathan, 1993; Zumbo, 1999), although it is rarely used in practice, perhaps because of the time commitment. Consequently, little published empirical evidence demonstrates the effects of purification on LR DIF detection (French & Maller, 2007) or any other DIF detection procedure.

French and Maller (2007) found that purification resulted in an 18% increase in power and a 20% decrease in Type I errors when only uniform DIF was evaluated, ignoring the influence of purification on the primary advantage of LR over MH (i.e., detection of nonuniform DIF). Moreover, methodological studies have implemented purification when comparing various DIF detection procedure (e.g., SIBTEST, MH), yet these same studies did not purify with LR (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). French and Maller recommend the evaluation of the influence of purification on DIF detection rates with LR. Purification should be included in future studies.

Second, only uniform DIF was investigated yet nonuniform DIF is a threat to validity. For example, Maller (2001) found that 16% of DIF items in a major standardized intelligence test were nonuniform. Finch and French (2007) examined previous nonuniform or crossing-DIF (CDIF) detection work by comparing the performance of four single-level procedures: LR, SIBTEST, IRTLR, and CFA. The researchers recognized that practitioners are interested in investigating both uniform and nonuniform DIF, and therefore, the current simulation study could be extended with the goal of detecting nonuniform DIF.

Third, as in most simulation studies, study conditions were limited due to time and computer capacity, thereby findings may not generalize to all possible conditions and may not be valid for other situations. For example, DIF was only simulated to the within-cluster condition, whereas the between cluster condition is often found in simulation literature on DIF. In addition, only two ICC values were used, .05 and .25, in contrast to other researchers who used a wider range.

Fourth, in this study, DIF was confined to level 1. In future research, a random effect for the item of interest across clusters could be introduced. If the random effect is significant, this implies that the item difficulty varies across clusters, meaning that the item functions differentially in different classrooms. If there is insignificant random effect of items, student-level DIF can be tested, followed by the detection of teacher-level DIF.

Last, a source of error in the MGMH analyses was the inability to efficiently and consistently analyze the simulated test sets with the SAS Begg adjustment method code, resulting in missing *p*-values. If this study is replicated, the source of this error must be definitively identified and corrected, and thereby capture all the *p*-values.

An extension to the multilevel DIF model used in this study is proposed for future research and involves the modeling of random effects at the between cluster level within the original model. When multilevel models are used, the hierarchical structure of data is considered. Moreover, the variance components are decomposed into the appropriate level so the homogeneity of students in a class or school can be modeled. However, the multilevel models considered herein only considered the intercept to be random and did not consider the slope to be random. By mirroring the approach of Kamata (2001) and

Kamata et al. (2005) with three-level Rasch model in which they allowed the coefficient of the person-level DIF to be random across higher-level clusters, such as schools, this model could be extended. This extension would be a random-effect DIF model, indicating that the effect of classroom or school membership can vary from unit to unit. The choice to parameterizing DIF as a random-effects DIF model rather than fixed effects and to estimate the variance of the DIF magnitude across higher-level clusters could be a useful extension to this research.

Appendix A

```
## This function takes person, cluster, and item parameters and generates a random response

item_response <- function(beta, nu, delta, transition_locations) {

 denominator <- 1 + exp(beta+nu-delta-transition_locations[1]) +

                exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2]) +

                exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2] + beta+nu-delta-transition_locations[3]) +

                exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2] + beta+nu-delta-transition_locations[3] + beta+nu-delta-transition_locations[4])

 prob_0 <- 1 / denominator

 prob_1 <- exp(beta+nu-delta-transition_locations[1]) / denominator

 prob_2 <- exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2]) / denominator

 prob_3 <- exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2] + beta+nu-delta-transition_locations[3]) / denominator

 prob_4 <- exp(beta+nu-delta-transition_locations[1] + beta+nu-delta-transition_locations[2] + beta+nu-delta-transition_locations[3] + beta+nu-delta-transition_locations[4]) / denominator

 p <- runif(1, 0, 1)

 if (p < prob_0) {

   0

 } else if (p < prob_0 + prob_1) {

   1

 } else if (p < prob_0 + prob_1 + prob_2) {

   2

 } else if (p < prob_0 + prob_1 + prob_2 + prob_3) {

   3

 } else if (p < prob_0 + prob_1 + prob_2 + prob_3 + prob_4) {
```

```
  4
 } else {
  5
 }
}
## To make this data generation replicable, set the seed (random.org)
set.seed(431124659)
## FIX CONSTANTS AND CREATE SIMULATION CONDITION MATRIX
{
# Fix item parameters (we will change items 19 and 20 to create DIF
transition_locations <- matrix(rep(c(-1.16,-.29,.32,1.13,-.89,-.33,.35,.87,-1.09,-
.69,.2,1.58,
                    -1.14,-.71,.22,1.64,-1.25,-.38,0.17,1.46,-1.54,-.30,.44,1.41,
                    -1.04,-.38,.28,1.13,-1.11,-.57,.10,1.58,-1.31,-.40,.27,1.44,
                    -1.29,-.40,.27,1.41),2),nrow=20,ncol=4,byrow=TRUE)
colnames(transition_locations) <- c("ti1","ti2","ti3","ti4")
item_difficulties <- rep(c(0.81, 1.07, 0.72, 0.58, 0.87, 0.93, 1.05, 0.88, 1.00, 0.93), 2)
# Simulation Details
num_reps        <- 400    # number of iterations per condition = 400
# Conditions
num_cluster     <- c(25, 50,100)
pers_per_cluster <- c(10, 20, 40)
dif             <- c(0, .2, .4, .8)
icc             <- c(.05, .25)
# Make a matrix of all conditions
conditions = expand.grid(num_cluster, pers_per_cluster, dif, icc)
colnames(conditions) <- c("num_cluster", "pers_per_cluster", "dif", "icc")
# Number of conditions so we can loop over them
num_conditions <- nrow(conditions)
```

```
}
## Loop over conditions
for (c in 1:num_conditions) {
  # DIF happens here.
  dif_item_diff <- item_difficulties
  dif_item_diff[19] <- dif_item_diff[19] + conditions$dif[c]
  dif_item_diff[20] <- dif_item_diff[20] + conditions$dif[c]
  ## Loop over replications
  for (iteration in 1:num_reps){
    # Draw cluster and person abilities
    beta <- rnorm(conditions$num_clust[c]*conditions$pers_per_cluster[c], 0, 1)
    nu   <- rnorm(conditions$num_clust[c], 0, conditions$icc[c]/(1-conditions$icc[c]))
    # Create empty dataframe to hold the data
    sim_data <- data.frame()
    # Loop over clusters then persons to cycle through all the persons
    for (clust in 1:conditions$num_cluster[c]) {
      for (pers in  1:conditions$pers_per_cluster[c]) {
        # Make a vector of person variables (id, cluster, group)
        pers_id <- (clust-1)*conditions$pers_per_cluster[c]+pers
        sim_pers <- c(pers_id, clust, pers %% 2)
        # Make vector of item responses - different item parameters if in group 0 or 1
        if (pers %% 2 == 0) { # Reference group has no DIF
          sim_items <- c()
          # loop over items, adding the responses
          for (i in 1:20) {
            sim_items <- c(sim_items, item_response(beta[pers_id], nu[clust],
item_difficulties[i], transition_locations[i, ]))
          }
        } else { # Focal group has DIF
```

```
        sim_items <- c()
        # loop over items, adding the responses
        for (i in 1:20) {
          sim_items <- c(sim_items, item_response(beta[pers_id], nu[clust],
dif_item_diff[i], transition_locations[i, ]))
        }
       }
       # Add to sim_data
       sim_data <- rbind(sim_data, c(sim_pers, sim_items))
     }
   }
   # Put names onto sim_data and save to file
   names(sim_data) <- c("PersID", "Clust", "Group", paste0("Item", 1:20))
   write.csv(sim_data, paste0("TestSet", (c-1)*num_reps+iteration, ".csv"), row.names =
FALSE)
  }
}
```

Appendix B

```
/* French, Finch, & Iverson, 2015

%macro pvals(data,item,pvalues);
/*Taking mean of items 1-20*/
/*Take mean of items 1-18 when analyzing data with DIF - purify the
total score*/
/*See French & Maller (2007)*/
data b1; set &data;
ability = mean (of item1-item20); /*20 items*/
run;

/*Assuming a cluster is Clust*/
data Clust;
   set b1;
keep Clust;
rename Clust=id;

/*Assuming this is for items*/
data &item;
   set b1;
keep &item;
rename &item=response;

/*Assuming exposure is Group - this is column 3 - and it is coded 1-0
for reference and focal*/
data Group;
   set b1;
keep Group;
rename Group=exposure;

data ability;
   set b1;
keep ability ;
rename ability=stratum;

data mhdat;
   merge Clust &item Group ability;

ods output covB=naivecovb;
proc glimmix empirical data=mhdat;
      model response = exposure / dist=normal
                                  covb;
run;

ods output covB=geecovb;
proc glimmix empirical data=mhdat;
      class id;
      model response = exposure / dist=normal
                                  covb;
      random _residual_ / subject=id type=cs vcorr;
run;

proc freq data=mhdat;
   table stratum*exposure*response / cmh;
   output out=mhresults cmh;
```

```
   run;

data naivecovb2;
   set naivecovb;
if _n_=1;
keep col1;
rename col1=naivevar;

data geecovb2;
   set geecovb;
if _n_=1;
keep col1 geevar;
rename col1=geevar;

data mhresults2;
   set mhresults;
keep _cmhcor_ _cmhrms_ _cmhga_;

data &pvalues;
format item $char7.;
item="&item";
   merge mhresults2 naivecovb2 geecovb2;
f=geevar/naivevar;
*bcmhcor=_cmhcor_/f;
*bcmhrms=_cmhrms_/f;
bcmhga=_cmhga_/f;
*bcmhcor85=_cmhcor_/(.85*f);
*bcmhrms85=_cmhrms_/(.85*f);
bcmhga85=_cmhga_/(.85*f);
*bcmhcor9=_cmhcor_/(.9*f);
*bcmhrms9=_cmhrms_/(.9*f);
*bcmhga9=_cmhga_/(.9*f);
*bcmhcor95=_cmhcor_/(.95*f);
*bcmhrms95=_cmhrms_/(.95*f);
*bcmhga95=_cmhga_/(.95*f);

*cmhcor_p=1-probchi(_cmhcor_,1);
*cmhrms_p=1-probchi(_cmhrms_,1);
cmhga_p=1-probchi(_cmhga_,4);
*bcmhcor_p=1-probchi(bcmhcor,1);
*bcmhrms_p=1-probchi(bcmhrms,1);
bcmhga_p=1-probchi(bcmhga,4);
*bcmhcor85_p=1-probchi(bcmhcor85,1);
*bcmhrms85_p=1-probchi(bcmhrms85,1);
bcmhga85_p=1-probchi(bcmhga85,4);
*bcmhcor9_p=1-probchi(bcmhcor9,1);
*bcmhrms9_p=1-probchi(bcmhrms9,1);
*bcmhga9_p=1-probchi(bcmhga9,4);
*bcmhcor95_p=1-probchi(bcmhcor95,1);
*bcmhrms95_p=1-probchi(bcmhrms95,1);
*bcmhga95_p=1-probchi(bcmhga95,4);

proc print;
run;
%mend;
```

```
/*Running macro*/
%macro maketable(number);
      %do i=1 %to &number;
            PROC IMPORT OUT= WORK.b1
                  /*CHANGE THE LINE BELOW TO THE FILE LOCATION OF TEST
SETS*/
                  DATAFILE=
"C:\Users\mdto223\Dropbox\Carol_Simulation\Checking Work\Checking SAS
Syntax\Data1\TestSet&i..csv"
            DBMS=csv REPLACE;
            RUN;
            %pvals(b1,item1,all1)
            %pvals(b1,item2,all2)
            %pvals(b1,item3,all3)
            %pvals(b1,item4,all4)
            %pvals(b1,item5,all5)
            %pvals(b1,item6,all6)
            %pvals(b1,item7,all7)
            %pvals(b1,item8,all8)
            %pvals(b1,item9,all9)
            %pvals(b1,item10,all10)
            %pvals(b1,item11,all11)
            %pvals(b1,item12,all12)
            %pvals(b1,item13,all13)
            %pvals(b1,item14,all14)
            %pvals(b1,item15,all15)
            %pvals(b1,item16,all16)
            %pvals(b1,item17,all17)
            %pvals(b1,item18,all18)
            %pvals(b1,item19,all19)
            %pvals(b1,item20,all20)
            /*combining pvalue tables together*/
            data allitems;
                  set all1 all2 all3 all4 all5 all6 all7 all8 all9
all10 all11 all12 all13 all14 all15 all16 all17 all18 all19 all20;
            proc print;
            run;
            /*CHANGE THE BELOW LINE TO THE FILE LOCATION OF THE OUTPUT
TABLES*/
            proc export data=allitems dbms=csv
outfile="C:\Users\mdto223\Dropbox\Carol_Simulation\Checking
Work\Checking SAS Syntax\OutputTable&i..csv" replace;
            run;
            /*dm 'log;clear;output;clear;'*/
      %end;
%mend;

%maketable(5)
```

Appendix C

```
#install.packages("ordinal")
library(ordinal)
library(parallel)
# Build up results matrix with conditions data
{
 # Conditions
 num_cluster      <- c(25, 50, 100)
 pers_per_cluster <- c(10, 20, 40)
 dif              <- c(0, .2, .4, .8)
 icc              <- c(.05, .25)
 # Make a matrix of all conditions
 conditions = expand.grid(num_cluster, pers_per_cluster, dif, icc)
 colnames(conditions) <- c("num_cluster", "pers_per_cluster", "dif", "icc")
 # Copy each row 400 times for the iterations.
 repconditions <- conditions[sort(as.numeric(rep(rownames(conditions), 400))),]
}
sim_analysis <- function(i) {
 temp_data <- read.csv(paste0("Data/TestSet", i, ".csv"))
 ## create a total score upon which we can regress item responses
 temp_data$Total <- rowSums(temp_data[,4:23])
 # outcomes need to be factors for clmm to use ordinal logistic regression
 # This line doesn't work, and it's really bothering me
 # temp_data[,4:23] <- apply(temp_data[,4:23], 2, as.factor)
 # So I'll do it the lazy way
 for (x in 4:23) {temp_data[,x] <- as.factor(temp_data[,x])}
 pvals <- NULL
 for (x in 4:23) {
```

```
   modelA <- clm(as.formula(paste0(colnames(temp_data)[x], " ~ Total")), data =
temp_data, Hess = FALSE)

   modelB <- clm(as.formula(paste0(colnames(temp_data)[x], " ~ Total + Group")), data
= temp_data, Hess = FALSE)

   #model3 <- clmm(as.formula(paste0(colnames(temp_data)[x+3], " ~ Total + Group +
Total*Group + (1 | Clust)")), data = temp_data)

   #pvals <- c(pvals, pchisq(2*(modelB$logLik-modelA$logLik), df = 1, ncp = 0,
FALSE))

   pvals <- c(pvals, anova(modelA, modelB)$`Pr(>Chisq)`[2])

  }

  c(repconditions[i,1], repconditions[i, 2], repconditions[i, 3], repconditions[i, 4], pvals)

}

# set up clusters

{

cl <- makeCluster(detectCores(logical = FALSE))

clusterEvalQ(cl, library(ordinal))

clusterEvalQ(cl, setwd("D:/Dropbox/4. Completed Projects/Carol_Simulation"))

clusterExport(cl, "repconditions")

clusterExport(cl, "sim_analysis")

}

results1 <- parLapply(cl, 1:24, sim_analysis)

results1 <- as.data.frame(do.call(rbind, results1))

# Let's give the columns reasonable names

colnames(results1)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")

colnames(results1)[5:24] <- c(paste0("p_item", 1:20))

# Save this data before it's too late!

write.csv(results1, "SimulationAnalysisResults1_NoRandom.csv", row.names = FALSE)

results2 <- parLapply(cl, 4001:8000, sim_analysis)

results2 <- as.data.frame(do.call(rbind, results2))

# Let's give the columns reasonable names
```

```
colnames(results2)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results2)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results2, "SimulationAnalysisResults2_NoRandom.csv", row.names = FALSE)
results3 <- parLapply(cl, 8001:12000, sim_analysis)
results3 <- as.data.frame(do.call(rbind, results3))
# Let's give the columns reasonable names
colnames(results3)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results3)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results3, "SimulationAnalysisResults3_NoRandom.csv", row.names = FALSE)
results4 <- parLapply(cl, 12001:16000, sim_analysis)
results4 <- as.data.frame(do.call(rbind, results4))
# Let's give the columns reasonable names
colnames(results4)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results4)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results4, "SimulationAnalysisResults4_NoRandom.csv", row.names = FALSE)
results5 <- parLapply(cl, 16001:20000, sim_analysis)
results5 <- as.data.frame(do.call(rbind, results5))
# Let's give the columns reasonable names
colnames(results5)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results5)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results5, "SimulationAnalysisResults5_NoRandom.csv", row.names = FALSE)
results6 <- parLapply(cl, 20001:24000, sim_analysis)
results6 <- as.data.frame(do.call(rbind, results6))
# Let's give the columns reasonable names
```

```
colnames(results6)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results6)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results6, "SimulationAnalysisResults6_NoRandom.csv", row.names = FALSE)
results7 <- parLapply(cl, 24001:28000, sim_analysis)
results7 <- as.data.frame(do.call(rbind, results7))
# Let's give the columns reasonable names
colnames(results7)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results7)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results7, "SimulationAnalysisResults7_NoRandom.csv", row.names = FALSE)
results8 <- parLapply(cl, 28001:28800, sim_analysis)
results8 <- as.data.frame(do.call(rbind, results8))
# Let's give the columns reasonable names
colnames(results8)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results8)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results8, "SimulationAnalysisResults8_NoRandom.csv", row.names = FALSE)
results <- rbind(results1, results2, results3, results4, results5, results6, results7, results8)
# Let's give the columns reasonable names
colnames(results)[1:4] <- c("num_cluster", "pers_per_cluster", "dif", "icc")
colnames(results)[5:24] <- c(paste0("p_item", 1:20))
# Save this data before it's too late!
write.csv(results, "SimulationAnalysisResults_NoRandom.csv", row.names = FALSE)
stopCluster(cl)
```

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. doi:10.1007/BF02293814

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley. doi:10.1002/0470114754

Atar, B. (2007). Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and Gllamm procedures. (Doctoral dissertation). Available from Florida State University's Digital Repository. (FSU_migr_etd-0248 (IID))

Begg, C. (1999). Analyzing *k* (2 x 2) tables under cluster sampling. *Biometrics*, 55, 302-307. doi:10.1111/j.0006-341X.1999.00302.x

Beaver, J. L., French, B. F., Finch, W. H., & Ullrich-French, S. C. (2014). Sex differential item functioning in the Inventory of Early Development III Social-Emotional Skills. *Journal of Psychoeducational Assessment*, *32*(8), 775-780. doi:10.1177/0734282914544924

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. doi:10.1007/BF02291411

Burkes, L. L. (2009). *Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom*

opportunities to learn. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3360218).

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Castle, C. (2018). *Measuring multidimensional science learning: Item design, scoring, and psychometric considerations* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (ProQuest No. 10786367)

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective* (pp. 3779-3779). (Doctoral dissertation). Available from Florida State University Libraries, Electronic Theses, Treatises and Dissertations.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333-353. doi:10.1111/j.1745-3984.1996.tb00496.x

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*(1), 57-79. doi:10.1207/s15327574ijt0601_4

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice*, *17*(1), 31-44. doi:10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., Kane, M. T., & Kim, S. H. (2001). The precision of simulation study results. *Applied Psychological Measurement*, *25*(2), 136-145. doi:10.1177/01466210122031966

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education*, *72,* 181-193. doi:10.3200/JEXE.72.3.181-193

Dodeen, H., & Johanson, G. (2001, April). The prevalence of gender-related DIF in survey data. Paper presented at the annual meeting of the American Educational Research Association Seattle, WA.

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* (Chapter 3, pp. 35-66). 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. doi:10.1111/j.1745-3984.1986.tb00255.x

Every Student Succeeds Act 2015. (2015). Retrieved from https://www.govinfo.gov/content/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Finch, W. H., Finch, M. E., & French, B. F. (2016). Recursive Partitioning to Identify Potential Causes of Differential Item Functioning in Cross-National Data. *International Journal of Testing*, *16*(1), 21-53. doi:10.1080/15305058.2015.1039644

Finch, W., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and psychological Measurement*, *67*(4), 565-582. doi: 10.1177/0013164406296975

Finch, H., & French, B. (2010). Detecting differential item functioning of a course satisfaction instrument in the presence of multilevel data. *Journal of the First-Year Experience & Students in Transition*, *22*(1), 27-47.

Francis, X. H. (2015). *Models for Hierarchical-Structured Item Response Data and a Longitudinal Multilevel Logistic Regression Model on DIF Analyses.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3738156)

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3), 299-317. doi:10.1111/j.1745-3984.2010.00115.x

French, B. F., & Finch, W. H. (2013). Extensions of Mantel–Haenszel for multilevel DIF detection. *Educational and Psychological Measurement*, *73*(4), 648-671. doi:10.1177/0013164412472341

French, B. F., & Finch, W. H. (2015). Transforming SIBTEST to account for multilevel data structures. *Journal of Educational Measurement*, *52*(2), 159-180. doi:10.1111/jedm.12071

French, B., Finch, W. H., & Imekus, J. (2019). Multilevel generalized Mantel-Haenszel for differential item functioning detection. *Frontiers in Education, 4(*47*),*1-10. doi: 103389/feduc.2019.00047

French, B.F., Finch, H., & Iverson, A.E.F. (2015). *Multilevel Differential Item Functioning Program, SAS version 1.0.1.* Available from the Learning and Performance Research Center, Washington State University.

French, B. F., Finch, W. H., & Vazquez, J. A. V. (2016). Differential Item Functioning on mathematics items using multilevel SIBTEST. *Psychological Test and Assessment Modeling*, *58*(3), 471.

French, B. F., Finch, W. H., Randel, B., Hand, B., & Gotch, C. M. (2016). Measurement invariance techniques to enhance measurement sensitivity. *International Journal of Quantitative Research in Education*, *3*(1-2), 79-93. doi:10.1504/IJQRE.2016.073672

Garrett, P. L. (2009). *A Monte Carlo study investigating missing data, differential item functioning, and effect size.* (Doctoral dissertation). ProQuest Dissertations and Theses database. (UMI Number: 3401601)

Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In *Handbook of multilevel analysis* (pp. 237-274). Springer, New York, NY. doi:10.1007/978-0-387-73186-5_6

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test Validity*, 129-145.

Holahan, Young, Palmer, & Little. (2017). Making the most of ESSA: 20 ideas for how to leverage ESSA to advance college and career readiness and equity. Education Counsel Retrieved http://educationcounsel.com/?publication=making-essa-twenty-questions-developing-implementing-strong-state-essa-plans-advance-college-career-readiness-equity

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi:10.1080/10705519909540118

Jin, Y., Myers, N. D., & Ahn, S. (2014). Complex versus simple modeling for DIF detection: when the intraclass correlation coefficient (ρ) of the studied item is less than the ρ of the total score. *Educational and Psychological Measurement*, *74*(1), 163-190. doi:10.1177/0013164413497572

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93. doi:10.1111/j.1745-3984.2001.tb01117.x

Kamata, A. & Binici, S. (2003). *Random Effect DIF Analysis via Hierarchical Generalized Linear Modeling*. Paper presented at the biannual International Meeting of the Psychometric Society, Sardinia, Italy.

Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (2005, April). Random-effect differential item functioning across group unites by the hierarchical generalized linear model. In *annual meeting of the American Educational Research Association, Montreal, Canada*.

Kim, J. (2018). *Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (ProQuest Number: 10930557)

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement*, *45*(3), 271-285. doi:/10.1111/j.1745-3984.2008.00064.x

Li, H., Qin, Q., & Lei, P. W. (2017). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment*, *22*(1), 1-17. Doi:10.1080/10627197.2016.1271702

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Abingdon, United Kingdom: Routledge Publications.

Massachusetts Department of Education. (2018). District and school accountability. 2018-MA-math. Retrieved from http://www.doe.mass.edu/accountability/lists-tools.html

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). New York: Springer. doi:10.1007/978-1-4757-2691-6_6

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*(2), 107-122. doi:10.1111/j.1745-3984.1993.tb01069.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30. doi:10.1002/j.2333-8504.1992.tb01436.x

Nolan, K. A. (2016). *Ignoring hierarchical data structure in item response theory analyses: Implications for educational and psychological research*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (ProQuest Number: 10583747)

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Educational and Psychological Measurement*, *72*(1), 44-51. doi:10.1177/0013164411409743

Park, H. (2008). *Comparison of IRT models for ordered polytomous response data*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9983591)

Peck, D. (2017). *Motivation to persist: The role of hope, academic self-efficacy, and sense of belonging on first generation Latinx college students and their intent to persist* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (ProQuest No. 10599064)

Penfield, R. D & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics*, *26* (pp. 125-167). Amsterdam: Elsevier. doi:10.1016/S0169-7161(06)26005-X

Porter, S. C., & Newman, L. (2016). A brief measure of attitudes toward resident advisors. *College Student Journal*, *50*(1), 107-111. doi:10.1037/t63763-000

Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377-399. doi:10.1177/0013164413507063

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

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. doi:10.1177/014662169001400208

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Reise, S. P, Ventura, J., Nuechterlein, K. H. & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, 84:2, 126-136. doi:10.1207/s15327752jpa8402_02

Ryan, C. H. (2008). *Using hierarchical generalized linear modeling for detection of differential item functioning in a polytomous item response theory framework: An evaluation and comparison with the generalized Mantel-Haenszel* (Doctoral dissertation). Available from https://scholarworks.gsu.edu/eps_diss/21/ (UMI No. 3323230)

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*. doi:10.1007/BF03372160

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M. Petersen, M., Sprangers, M., & EORTC Quality of Life Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of clinical epidemiology*, *62*(3), 288-295. doi: 10.1016/j.jclinepi.2008.06.003

Sharafi, Z., Mousavi, A., Ayatollahi, S. M. T., & Jafari, P. (2017). Assessment of differential item functioning in health-related outcomes: A simulation and empirical analysis with hierarchical polytomous data. *Computational and Mathematical Methods in Medicine*, *2017*. doi:10.1155/2017/7571901

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. doi:10.1007/BF02294572

Shear, B. R. (2018). Using hierarchical logistic regression to study DIF and DIF variance in multilevel data. *Journal of Educational Measurement*, *55*(4), 513-542.

Snijder, T. A. & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313-350. doi: 10.1207/s15324818ame1804_1

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

Swanson, D. B., Clauser, B. E., Case, S. M., Nungster, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53–75. doi:10.3102/10769986027001053

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3-46. doi:10.1177/1094428114553062

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577. doi:10.1007/BF02295596

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*(2), 272-296. doi:10.1177/0013164404268667

U.S. Department of Education. (2017). Every Student Succeeds Act. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/essaassessmentfactsheet1207.pdf

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach* (Doctoral dissertation). Available from the FSU Digital Library.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. doi:10.1177/0734282911406666

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166-180. doi:10.1177/0146621609355279

Wang, W. C., & Su, Y. H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*(6), 450-480. doi: 10.1177/0146621604269792

Wen, Y. (2014). *DIF analyses in multilevel data: Identification and effects on ability estimates* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3665404)

Williams, N. J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous item response theory models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3116234)

Wilson, J. R., & Lorenz, K. A. (2015). *Modeling binary correlated responses using SAS, SPSS and R* (Vol. 9). Springer. doi:10.1007/978-3-319-23805-0

Wood, W. S. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3461258)

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251. doi:10.1111/j.1745-3984.1993.tb00425.x

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Describing and categorizing DIF in polytomous items. *ETS Research Report Series*, *1997*(1), i-52. doi:10.1002/j.2333-8504.1997.tb01726.x

Vita

Carol D. Hanley, Ed. D

PROFESSIONAL PREPARATION

- Post-Baccalaureate, evaluation, University of Kentucky, Lexington, KY, 2013 - 2016
- Ed.D., Curriculum and Instruction, 1997
  - University of Kentucky, Lexington, KY
  - Dissertation: *The effects of the learning cycle on the ecological knowledge of general biology students as measured by two assessment techniques.*
- Rank I, Biology, 1986
  - Eastern Kentucky University, Richmond, KY
- M.S., Environmental Studies, 1980
  - State University of New York at Buffalo, Buffalo, NY
  - Thesis: *The use of silica-geothermometers to determine geothermal gradients in Central and Western New York.*
- B.A., Biology, 1978
  - State University of New York at Buffalo, Buffalo, NY

APPOINTMENTS

- **Office for Environmental Outreach Services, College of Agriculture, Food and Environment**, University of Kentucky, Lex, KY, 2013 – present
- **International Programs, College of Agriculture, Food and Environment**, University of Kentucky, Lex, KY, 2013 – present
- **Research Office, College of Agriculture, Food and Environment**, University of Kentucky, Lex, KY, 2009 – present
- **Environmental and Natural Resources Initiative, College of Agriculture, Food and Environment**, associate director, University of Kentucky, Lex, KY, 2009 – 2013
- **Associate Director, Tracy Farmer Institute for Sustainability and the Environment (TFISE)** (formerly Director of Education and Communications, Tracy Farmer Center for the Environment (TFCE)), University of Kentucky, Lex., KY, 2005 – present
- **4- H Youth Development Extension Specialist and Associate Director of P-12 Environmental Education**, Tracy Farmer Center for the Environment, University of Kentucky, Lex., KY, 2001 – 2005

- **Sciences Branch Manager, Kentucky Department of Education,** 1997 - 2001
- **Science Teacher, Biology**, Bryan Station High School, Lexington, KY, 1984 - 1997
- **Environmental Consultant, Ecology and Environment Inc**., Buffalo, NY, 1980 - 1983

Publications

- Hanley, C. & Taylor, K. (2017). Water as the context for community-based science projects: Teaching the next generation. In B. D. Lee, D. I. Carey, & A. L. Jones (Eds.), *Water in Kentucky: Natural history, communities, and conservation* (213-222). Lexington, KY: University Press of Kentucky.
- Hanley, C., Davis H., & Davey, B. (2012). The impact of professional development in natural resource investigations using geospatial technologies. *Journal of Natural Resources and Life Science Education*, 41, 68–78.
- 4-H Curriculum Development: The Power of the Wind, Butterfly WINGS, Food, Culture and Reading, Exploring the Environment, Forestry