Spring 5-12-2023

A Comparison Study of the Differential Functioning of Tests Statistic and a New Mahalanobis Distance-Based Statistic For Pre-Screening Item Response Theory Models

David Fikis

ACCEPTANCE

This dissertation, A COMPARISON STUDY OF THE DIFFERENTIAL FUNCTIONING OF TESTS STATISTIC AND A NEW MAHALANOBIS DISTANCE-BASED STATISTIC FOR PRE-SCREENING ITEM RESPONSE THEORY MODELS, by DAVID ROBERT JOHN FIKIS, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education & Human Development, Georgia State University. The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

_____
Hongli Li, Ph.D.
Committee Chair

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
Audrey Leroux, Ph.D.
Committee Member

_____
Yinying Wang, Ph.D.
Committee Member

_____
Date

_____
Jennifer Esposito, Ph.D.
Chairperson, Department of Educational Policy Studies

_____
Paul A. Alberto, Ph.D.
Dean, College of Education & Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education & Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

DAVID ROBERT JOHN FIKIS

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

David Robert John Fikis
Educational Policy Studies
College of Education & Human Development
Georgia State University

The director of this dissertation is:

Dr. Hongli Li
Department of Educational Policy Studies
College of Education & Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

David Robert John Fikis

ADDRESS: 30 Pryor St SW # 450
Atlanta, GA 30303-3219

EDUCATION:

Ph.D.	2023	Georgia State University Educational Policy Studies
M.S.	2011	Georgia State University Instructional Technology
B.A.	2008	Berry College Philosophy and Religion

PROFESSIONAL EXPERIENCE:

2011-present	Research Associate Georgia State University, Atlanta, GA
2019-2021	Center Manager Code Ninjas Atlanta
2010-2011	Graduate Administrative Assistant Georgia State University, Atlanta, GA

PRESENTATIONS AND PUBLICATIONS:

Leroux, A.J., Cappelli, C.J. and Fikis, D.R.J. (2021). The impacts of ignoring individual mobility across clusters in estimating a piecewise growth model. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12229

Fikis, D. R. J., & Oshima, T. C. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement, 77*(3), 415-428. doi:10.1177/0013164416645844

Wang, Y. Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational Administration Quarterly, 53*(2), 289-323. doi:10.1177/0013161x16660585

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2011-2018	American Education Research Association

A COMPARISON STUDY OF THE DIFFERENTIAL FUNCTIONING OF TESTS STATISTIC AND A NEW MAHALANOBIS DISTANCE-BASED STATISTIC FOR PRE-SCREENING ITEM RESPONSE THEORY MODELS

by

DAVID ROBERT JOHN FIKIS

Under the Direction of Hongli Li, Ph.D.

ABSTRACT

The Differential Test Functioning (DTF) statistic, with the Item Parameter Replication (IPR) procedure, can measure Differential Item Functioning (DIF) within the Differential Functioning of Items and Tests (DFIT) framework for Item Response Theory (IRT) models. However, it comes with many practical costs and theoretical assumptions. In some reasonably anticipated circumstances, the DTF statistic cannot be evaluated easily, and DFIT analysis consequentially remains beyond the scope of impacted IRT models. A straightforward, diagnostic statistic would add value to typical IRT model fitting. It was hypothesized that a statistic based on Mahalanobis distances and standard errors of an IRT model could perform as a reliable flag for likely DIF. To test this hypothesis, a Monte Carlo simulation study compared the performance of the traditional DTF measure to the new statistic. Although easy to calculate, the statistic proved unproductive in flagging models with DIF present. Related performance analysis and recommendations were provided.

INDEX WORDS: Item Response Theory, Differential Item Functioning, Monte Carlo Simulation, High-Performance Computing

A COMPARISON STUDY OF THE DIFFERENTIAL FUNCTIONING OF TESTS STATISTIC AND A NEW MAHALANOBIS DISTANCE-BASED STATISTIC FOR PRE-SCREENING ITEM RESPONSE THEORY MODELS

by

DAVID ROBERT JOHN FIKIS

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Research, Measurement, & Statistics

in

Educational Policy Studies

in

the College of Education & Human Development

Georgia State University

Atlanta, GA

2023

DEDICATION

And Raphael now, to Adam's doubt proposed,

Benevolent and facile thus replied.

To ask or search, I blame thee not; for Heaven

Is as the book of God before thee set,

Wherein to read his wondrous works, and learn

His seasons, hours, or days, or months, or years:

This to attain, whether Heaven move or Earth,

Imports not, if thou reckon right; the rest

From man or angel the great Architect

Did wisely to conceal, and not divulge

His secrets to be scanned by them who ought

Rather admire; or, if they list to try

Conjecture, he his fabric of the Heavens

Hath left to their disputes, perhaps to move

His laughter at their quaint opinions wide

— Milton

ACKNOWLEDGMENTS

We acknowledge the use of Advanced Research Computing Technology and Innovation Core (ARCTIC) resources at Georgia State University's Research Solutions made available by the National Science Foundation Major Research Instrumentation (MRI) grant number CNS-1920024.

List of Tables v

List of Figures vi

1 THE PROBLEM 1

Introduction 1

Problem Statement 5

Purpose of Study 6

Significance 7

Limitations 7

2 REVIEW OF THE LITERATURE 9

Item Response Theory Models 9

Item Response Theory Performance 15

Differential Item Functioning (DIF) 23

Differential Test Functioning (DTF) 34

Mahalanobis Distance 36

3 METHODOLOGY 40

Overview 40

Design 42

Procedures 44

Data Cleaning 48

4 RESULTS 54

Non-Convergence Errors 54

Overall Findings 55

Parameter Estimation 63

DIF Measurement 68

Computation Time 80

5 DISCUSSION 83

Comparison of DIF Measurements 83

Outcome Interpretations 84

Implications 90

Limitations and Future Research 92

Conclusions 98

REFERENCES 100

APPENDICES 121

List of Tables

Table 1 17

Table 2 43

Table 3 45

Table 4 50

Table 5 52

Table 6. 54

Table 7. 57

Table 8. 58

Table 9. 64

Table 10 65

Table 11 66

Table 12 67

Table 13. 67

Table 14. 68

Table 15. 69

Table 16 72

Table 17. 74

Table 18. 76

Table 19 77

Table 20. 80

Table 21. 81

List of Figures

Figure 1 2

Figure 2 3

Figure 3 17

Figure 4 24

Figure 5 25

Figure 6 26

Figure 7 34

Figure 8. 50

Figure 9 53

Figure 10 71

Figure 11 72

Figure 12 78

Figure 13 79

Figure 14 79

Figure 15 82

1 THE PROBLEM

Introduction

Item Response Theory (IRT) is an approach to measurement based on the assumption that a trait of interest is indirectly measurable by calculating the likelihood of a set of persons' responses to items with quantifiable parameters (Embretson & Reise, 2000; Hambleton et al., 2010). Differential Item Functioning (DIF) is a phenomenon within Item Response Theory where those likelihoods fail to remain consistent between all examinees, usually because of the interaction of an extraneous trait with some imperfect items (Oshima & Morris, 2008). A certain technical report puts it in clear language: "it means that examinees with identical θs will have different chances of getting the item correct (P(θ)), depending on their group. That situation is clearly unfair" (Warm, 1978, p. 128).

Within IRT, the evaluation of DIF is an evolving subject matter. DIF analysis methods carry computational and structural costs; the traditional hypothesis-based testing for DIF is complex. Some relatively unexplored avenues of DIF analysis, when applied in a manner informed by existing literature, may present a new, simpler option. Examining the *measurement failures* of a model based on the Mahalanobis distance might provide a relatively affordable "check engine light" for everyday practitioners and users of IRT to make data-informed decisions on whether to invest in using or testing a model further.

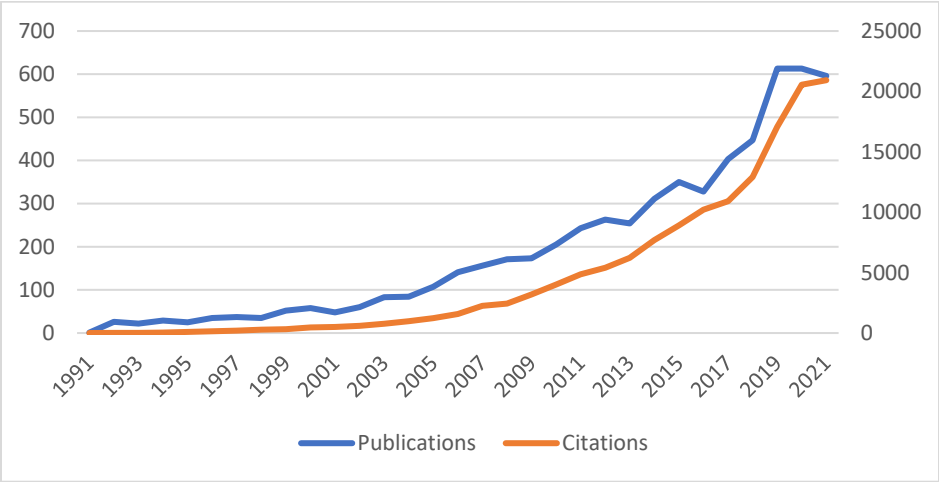
The Growth of IRT Scholarship

Both IRT and the tools for using it are well-established and continue to develop. IRT scholarship, going by its presence in the literature, ages well for an octogenarian: the number of publications and the citations they generate are strongly correlated, $r(28) = .98$, $p < .01$, when the

exponential growth is transformed logarithmically to satisfy the linearity assumption of the Pearson's *r*:

Figure 1

Publications and Citations, via Clarivate Analytics, for "Item Response Theory"



In practice, as well, IRT proliferates. It is stated that "all major educational tests…are developed using item response theory" (An & Yung, 2014, p. 1) Within the field of language testing, for example, the Rasch model has been found to be held in "wide acceptance" (McNamara & Knoch, 2012, p. 556). There are more R packages available for Item Response Theory than any other common theoretical framework in psychometrics (*CRAN Task View: Psychometric Models and Methods*, 2022).

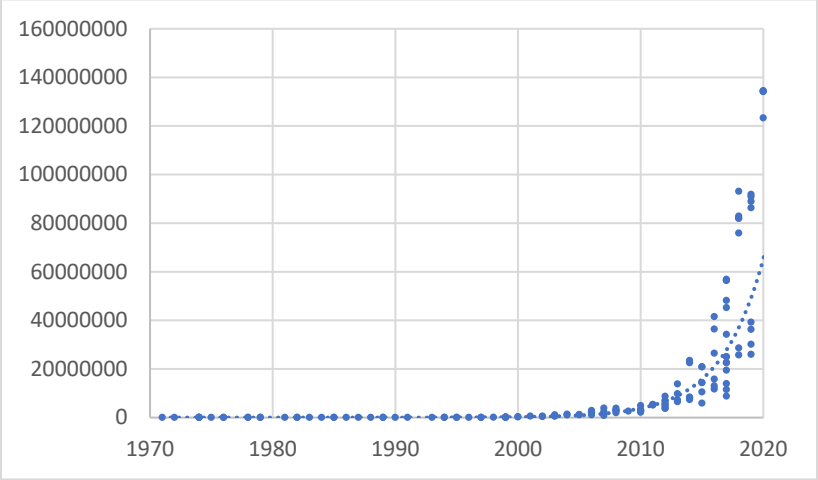
The Development of Computational Capacity

Similarly, microprocessor capability, going by achievements in production, proceeds apace: the density of transistors relative to the passage of time—Moore's Law—are strongly correlated, $r(240) = .99$, $p < .01$, with the same transformation:

Figure 2

CPU Transistors per mm^2, data via

https://gist.github.com/emartin59/0345adc1a60ad58433bb9b24113f490b



The development looks like it has never been better in both instances, but the development is not as sustainable as it may seem. Defenders of Moore's law tend to point out that continued gains in microprocessor engineering are only possible through multi-faceted innovations in diverse aspects of the process (Bohr, 2018; O'Boyle, 2019; Rotman, 2020). In research sectors, these hardware innovations are accessed after appropriate changes are made to software to capitalize on them (Fikis & Oshima, 2017; Putz et al., 2013; Sheng et al., 2014; Wang et al., 2017). However, even popular IRT programs show opportunities directly related to computational speed and effectiveness (Fikis & Oshima, 2017).

Even the physical tools—the computer hardware—used in IRT has an uncertain accessibility. Decreases in supply and increases in computer cost have been reported worldwide (Asianet-Pakistan, 2021; Garcia, 2020; Leesa-Nguansuk, 2021, 2022; Reporter, 2021). Reportedly, in the Philippines, the increased demand for computers in education has even inspired a call for government regulation of their price (Asianet-Pakistan, 2020). Economic

forecasts call for growth in GPU sectors, such as those used in high-performance computing, but increased profits might not translate to increased affordability (M2PressWIRE, 2020).

Thus, the growth of computational potential does not tell the full story of the efforts involved and rising concerns over the affordability of psychometric research and evaluation. Models emphasizing parsimony would present tangible advantages and may even make IRT accessible to groups that would not otherwise be able to afford the analysis.

The Nature of School and District Size

In winter of 2021, Georgia reported 1,572 Milestones scores aggregated by school and subject with a median number tested of 69 (GaDOE, 2021). Nationwide, the 17,521 public school districts which remained open throughout 2021 reported a median grade size of 56 students (*Common Core of Data (CCD)*, 2021).

Research has generally found that smaller class sizes present various, if contextual, desirable effects at multiple educational levels (Ake-Little et al., 2020; Bowne et al., 2017; Canbeldek & Isikoglu Erdogan, 2017; De Paola et al., 2013; Laitsch et al., 2021; Li & Konstantopoulos, 2016, 2017; Shen & Konstantopoulos, 2021). Some studies have claimed that leadership practices do not vary, "systematically," with district size (Burkman et al., 2019). More focused inquiry into class size in online environments has also found results suggesting complex relationships (Bowne et al., 2017; Lin et al., 2019; Lowenthal et al., 2019; Sorensen, 2015).

Thus, a great majority of environments where practical psychometric research could inform decisions are so small—with literature encouraging that smallness—that IRT may not be possible. Any approach to DIF detection with a reduced sample size burden therefore has an immediate relevance for educational policy; data-based statements about test fairness might be more possible in more diverse environments.

The Policymaking Implications of Theory, Practice, and Potential

From the perspective of the study of educational policy, then, the landscape for today's educational decisionmakers is an intersection of issues and constraints. The need for fair and reliable measurement remains a matter not far removed from the media a stakeholder may encounter, and continually evolving factors sustain the relevance of questions about what measurements are best to implement for the evaluation of education. DIF analysis, though well-established, may be challenging to implement because of costs and limits from smaller sample sized and/or smaller-sized populations from which to draw those samples. If IRT—and the DIF analyses to determine if it is being implemented with fairness and equity—were made more possible in those circumstances, they would be more effective within their appropriate contexts.

Problem Statement

Within IRT, DIF is essential but also expensive. In an examination of fairness in testing within the context of IRT, Bialo (2021, p. 8) astutely synthesizes policies from the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education by observing that, "a fair test is one in which scores have the same meaning for all test-takers." Fortunately, one of the primary assumptions of Item Response theory is measurement invariance; in an IRT model, the latent trait measured by an instrument is on the same scale and is a functions of the same underlying ability construct regardless of who takes the test (Hambleton et al., 2010). DIF is the term for when this fails to be true, and tests for it have been in development alongside IRT itself (Oshima & Morris, 2008).

The traditional, hypothesis-based testing used for DIF testing both at the item and test level presents a host of theoretical challenges including additional sample size requirements, specification of group membership, and constructing models under null and alternative

hypotheses (Cuhadar et al., 2021; De Boeck & Cho, 2021; Kopf et al., 2014b; O'Neill et al., 2020; Sahin & Anil, 2017; Schulze et al., 2022). These theoretical challenges increase practical burdens on the computational resources needed to test if a test is fair under the IRT framework with DIF analysis (Fikis & Oshima, 2017; Robitzsch & Ludtke, 2022). While alternative approaches that avoid some of these challenges and burdens are under continual development, there still remain some—such as using the Mahalanobis distance—that have not received in-depth analysis (Dimitrov, 2017; Lord, 1980; Penny, 1994; Randall & Engelhard, 2010). Considering the apparent pace DIF models outpace the growth of classroom sizes and accessible computer abilities, there is little reason to overlook the potential use of the Mahalanobis distance to aid the researcher in determining how badly they may be needed for a particular IRT model.

The primary theoretical limitation of existing DIF analysis methods is that a researcher must already have some notion of the problem; the need for a hypothesis to test in most DIF frameworks creates a certain need for *a priori* knowledge of problematic instrumentation. The primary practical limitation of existing DIF analysis methods is that a researcher must have an even larger sample size; certain approaches are simply beyond the scope of possibility for many potential decision-makers. In addition to these limitations, DIF tests also present non-trivial costs in time and resources. If a method existed by which a model could be characterized as *warranting* a DIF analysis rather than specifically *demonstrating* DIF, and if that method was relatively simple to execute and interpret, then navigating the evaluation process would be more straightforward and, perhaps, affordably reliable for practitioners.

Purpose of Study

The purpose of this study was to investigate a new statistic for detecting DIF in IRT models. It was based on the skewness of the distribution of Mahalanobis distances of the

standard errors of estimated item parameters. A simple procedure based on cut-off values empirically derived from comparisons with the DTF statistic could accompany reports of an IRT model fitting to provide an inexpensive "check engine light" for encouraging a researcher to engage in a more expensive DIF analysis. The research question was "How well do Mahalanobis distances perform as DIF indicators?"

To answer this question, this study proposed a simulation to investigate the effectiveness of using the skewness of the distribution of Mahalanobis distances of the standard errors of estimated item parameters as a "check engine light" to flag models that warrant further DIF analysis.

Significance

Mechanisms of Item Response Theory and Differential Item Functioning provide a well-researched and ever-improving theoretical framework for answering questions about test validity and fairness, but the cost of these methods is beyond the scope of many environments where important, formative educational decisions might be made. There is a great deal of good to be said about DIF as a means of investigating the fairness of a test, but policymakers may not be able to listen to the necessary explanations. A single, simple statistic with broad generalizability and low computational cost could bridge the gap between everyday measurement and more focused evaluation even if it came with footnotes.

Limitations

The primary limitation of the proposed statistic is a consequence of the method of calculation: because the Mahalanobis distance is used without a hypothesis testing framework, the interpretation of findings should be considered limited and perhaps even descriptive in nature. Just as a "check engine" light would not explain whether the problem is an oil leak,

missing catalytic converter, or otherwise, so too would this new statistic only be able to implicate rather than demonstrate DIF: it is simple, but it is not conclusive. It is an encouragement rather than a replacement of the DTF statistic and DFIT (differential functioning of items and tests) procedures in general. The justification and effectiveness of a hypothesis-based test for the proposed statistic remains a topic suitable for future research.

A theoretical limitation of the proposed study arises from assumptions of normality necessary to its methods. While the Central Limit Theorem establishes that sampling means are normally distributed regardless of the underlying distribution of their samples, parameter estimates are not themselves sampling means. Monte Carlo simulation methods provide ways to avoid this challenge by examining the mean performance of multiple repetitions to satisfy the Central Limit Theorem's requirements. Nevertheless, the distribution of item parameters in an instrument is not automatically the product of a sampling process, and strictly speaking, the standard errors calculated from the Hessian matrix could be argued to lack local independence since they are jointly estimated in a multivariate space. These theoretical issues may undermine the rationale in this and other related studies. What can be made to work in a simulation may not automatically generalize to real-world contexts. Examining those real-world models in detail remains a potentially meaningful—if not fruitful—topic for further research.

An additional limitation of the study comes from the nature of Monte Carlo simulations: although conditions were manipulated based on existing literature, conditions themselves were limited. Other conditions may be of interest to other researchers such as different levels of sample size or entirely new conditions such multiple group membership remain topics suitable for future research.

2 REVIEW OF THE LITERATURE

The purpose of this study was to evaluate the performance of a computationally simple statistic as an indicator of potential DIF. This statistic was based on the skewness of Mahalanobis distances of standard errors of parameter estimates. Unlike traditional test-based DIF analysis methods, this statistic needs neither the theoretical framework of an alternative hypothesis nor the practical framework of larger samples and multiple models to test. Thus, this statistic might add value as a routine "check engine light" to display during practical model fitting in Item Response Theory (IRT).

Accordingly, this study is set both within technical and theoretical aspects of specific applications of IRT which warrant a review of some relevant literature. First, an overview of the conceptual framework of the study—or, perhaps more astutely, an introduction to its mathematical context of IRT—is provided for the benefit of readers for whom the minutiae of IRT estimation algorithms are not ready-to-hand. Second, through a specific examination of the performance of IRT models, properties of interest are enumerated. Next, the concepts of DIF and the measurement of Differential Test Functioning (DTF) are provided. Finally, the Mahalanobis distance and Monte Carlo simulation are described to introduce the study's nature and method.

Item Response Theory Models

Gödel's Incompleteness Theorem posits that no mathematical system can be both *sound* (free of contradiction) and *complete* (capable of representing everything) and has been stretched outside its original context as an apt metaphor for the limits of complex systems (Hofstadter, 1979). It would be fair to apply this philosophical exercise to Item Response Theory as a limit to what any model might be capable of, but it is perhaps more comforting to the researcher to

metaphorically apply it as a limit to representations of the math; there is no approach to categorizing IRT models without limitations, whether from contradiction or lack of coverage.

Item Response Theory is a fecund theoretical framework. Efforts to classify IRT models display symptoms of that growth. For example, Hambleton et al. (2010) stated that "only a few models are in current use" (p. 12) and enumerated them with the following categories:

1. One-Parameter Logistic Model
2. Two-Paremeter Logistic Model
3. Three-Parameter Logistic Model
4. Other Promising Models

while shortly after, Linden (1996) produced a well-organized collection of over 100 models and sorted them in the following categories:

1. Common Models
2. Models for items with polytomous response formats
3. Models for response time or multiple attempts on items
4. Models for multiple abilities or cognitive components
5. Nonparametric models
6. Models for nonmonotone items
7. Models with special assumptions about the response process

where the common models are "the original unidimensional normal-ogive and logistic IRT models for items with dichotomously-scored responses" (p. vi). Later, Embretson and Reise (2000) offered the following for an introduction:

1. Binary IRT Models
    a. Unidimensional models

i. Traditional logistic models

ii. Traditional Normal Ogive Models

iii. Other Unidimensional Models

1. Models with restrictions on the parameter structure
2. Models for combining speed and accuracy
3. Single items with multiple attempts
4. Models with special forms for ICC

b. Multidimensional models

i. Exploratory multidimensional models for binary data

1. Multidimensional logistic models
2. Normal Ogive Models

ii. Confirmatory multidimensional models

1. Models for noncompensatory dimensions
2. Models for learning and change
3. Models with specified trait level structures
4. Models for distinct classes of persons

2. Polytomous IRT Models

a. The Graded-Response Model

b. The Modified Graded Response Model

c. The Partial Credit Model

d. The Generalized Partial Credit Model

e. Rating Scale Model

f. The Nominal Response Model

g. Continuous Response Model

and even still observed that "due to space considerations, several potentially important models are not described" (p. 101).

The 1-Parameter Logistic Model

Textbooks that introduce IRT models start with a unidimensional and dichotomous model—the 1-Parameter Logistic (1PL) model: pass/fail items measuring one thing—such as in Embretson and Reise (2000):

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{exp(\theta_s - \beta_i)}{exp(\theta_s - \beta_i)} \quad (1)$$

where "the simple probability that person s passes item *i*," *P*, depends on a personal trait θ_s and item difficulty β_i as well as in Hambleton et al. (2010, p. 12):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}; i = 1,2, \dots, n \quad (2)$$

where "a randomly chosen examinee with ability θ answers item i correctly" given item difficulty *b* with probability P. Both attribute the model to seminal work in Rasch (1960, p. 168):

$$\theta_{vi} = \frac{\xi_v}{\xi_v + \theta_i} \quad (3)$$

where the "probability (θ_{vi}) that a person (v) gives a correct answer to an item (i)" is described in the middle of a much more profound discussion on measurement in mathematical psychology, but the latter does implicitly acknowledge the multi-threaded beginnings to the model by briefly citing the work of Lord (1952), where students who memorize Rasch's 1PL model as origination *ex nihilo* might be surprised to see an "Item Characteristic Curve" (p. 7) and consideration of a "Discrimination Index" (p. 27) ahead of when they may otherwise anticipate such encounters. The parameterization—or labelling—of such related concepts tells more of the model than the math itself; for example, with a little effort, as illustrated by Templin (2008):

$$\begin{cases} \eta_{ij} = \pi_{0i} + \varepsilon_{ij} \\ \pi_{0i} = \beta_{0i} + \theta_{0j} \end{cases} \quad (4)$$

where notation based on Raudenbush and Bryk (2002, pp. 365-368) describes the probability of correct item response, η_{ij}, as a hierarchical model of item difficulty β with individual ability θ, most of the features of the 1PL model are replicated in a different approach with its own merits.

The 2-Parameter Logistic (2PL) Model

Compared to the 1PL, IRT models developed since its inception offer a more developed theoretical framework for researchers. A two-parameter model for dichotomous responses is described in Lord et al. (1968, p. 400) as:

$$P_g(\theta) = \Psi\left[1.7a_g\left(\theta - b_g\right)\right] \equiv \left(1 + e^{-Da_g\left(\theta - b_g\right)}\right)^{-1} \quad (5)$$

where Ψ is the logistic cumulative distribution function, θ is an ability, D is a scaling factor to bring the logistic model into equivalency with the normal model when set to 1.7, and a and b are item parameters such that b behaves similarly to item difficulty in Equation 2 and a is "discriminating power" manipulating the slope of the item characteristic curve—or, in another manner of speaking, the variance of the underlying distribution of responses (Lord & Novick, 1968, p. 367). Although useful and theoretically sound, manipulating the distribution in this way creates a plethora of mathematical challenges that can scarcely be understated.

The primary challenge can be appreciated in mathematical terms by examining the concept of statistic sufficiency. While discussing the normally-distributed nature of maximum likelihood estimators, Hambleton and Swaminathan (1984) observe that sufficient statistics exist for the 1PL and 2PL, but not for the 3PL nor "in any case of the normal ogive models" (p. 89). A sufficient statistic is one that can be used to replicate a sample or adequately determine an unknown parameter without using the original, random sample, such as the mean of a normally-

distributed population with a known variance or the maximum of a uniformly-distributed population (Kennedy, 2006).

The 3-Parameter Model (3PL)

The nature of multiple-choice tests introduces the possibility of guessing, and the 3-Parameter (3PL) model accounts for this guessing by creating a lower asymptote for the probability of a correct response. Hambleton et al. (2010, p. 17) describe it as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad (6)$$

with C representing a "pseudo-chance" parameter; they further cite Lord (1984) to caution against considering this as strictly a guessing parameter, but the cognomen has remained in use. Cuhadar et al. (2021) found that omitting pseudo-guessing parameters had varying but observable effects on estimation.

Other Parameterization

Other models exist within the unidimensional, dichotomous framework. For example, a 4-Parameter model produced an upper asymptote, such that:

$$P(\theta) = c + (\delta - c)F(\theta) \quad (7)$$

with the intention of minimizing the impact of response entry errors and tested using the SAT, GRE, and AP Calculus AB exams, but found "no compelling reason to urge the use of this model" (Barton & Lord, 1981, p. 6) Even the 4PL has been revisited and found to have some merits in the context of estimation methods not examined in its initial assessment (Culpepper, 2016) .

The first, theoretical polytomous models for IRT were published in the late 1960's around the same time as other early unidimensional variants, but more practical versions emerged 20-30 years later (Nering & Ostini, 2010, p. 24). The earliest formulations of the

Graded Response Model can be found in the thorough work of Samejima (1968) with discussion of both the theoretical model and detailed mathematical examination of estimators; in this original formulation, a unidimensional, latent trait is analyzed with free response items that are "classified into a certain limited number of categories arranged in the order of attainment or intensity" such that:

$$P_v(\theta) = \prod_{k_g \in V} P_{kg}(\theta) \quad (8)$$

where the probability of response pattern V is a function of ability θ, equal to the product of the locally independent probabilities of each responses k to every item g (pp. 3-4). A sample size of at least 500 responses is recommended to fit Graded Response Models (Embretson & Reise, 2000; Reise & Yu, 1990). For smaller samples (75 or 150) some improvements in parameter recovery are observed when using Markov Chain Monte Carlo estimation instead of Marginal Maximum Likelihood techniques (Kieftenbeld & Natesan, 2012).

Item Response Theory Performance

Much research is conducted to investigate the effectiveness of these related models under a wide variety of conditions. For example, within the context of a test of vocabulary, Holster and Lake (2016) investigated the misspecification criticisms Stewart (2014) made of Beglar (2010) using a 1PL Rasch model instead of a 3PL and found mixed results that, overall, encourage further investigation. Furthermore, Crocker (1986) pointed out that the 3PL has a disadvantage of requiring large sample sizes, and, later, Finch (2005) noted that even 1,000 individuals per group in a 20-item test had difficulty with estimation in certain contexts (Cuhadar et al., 2021). Kim and Lee (2017) simulated three-parameter models with sample size conditions of 500 and 3,000 while investigating the performance of item calibration methods for Bayesian estimation procedures and found that the benefits of various options were more observable in the

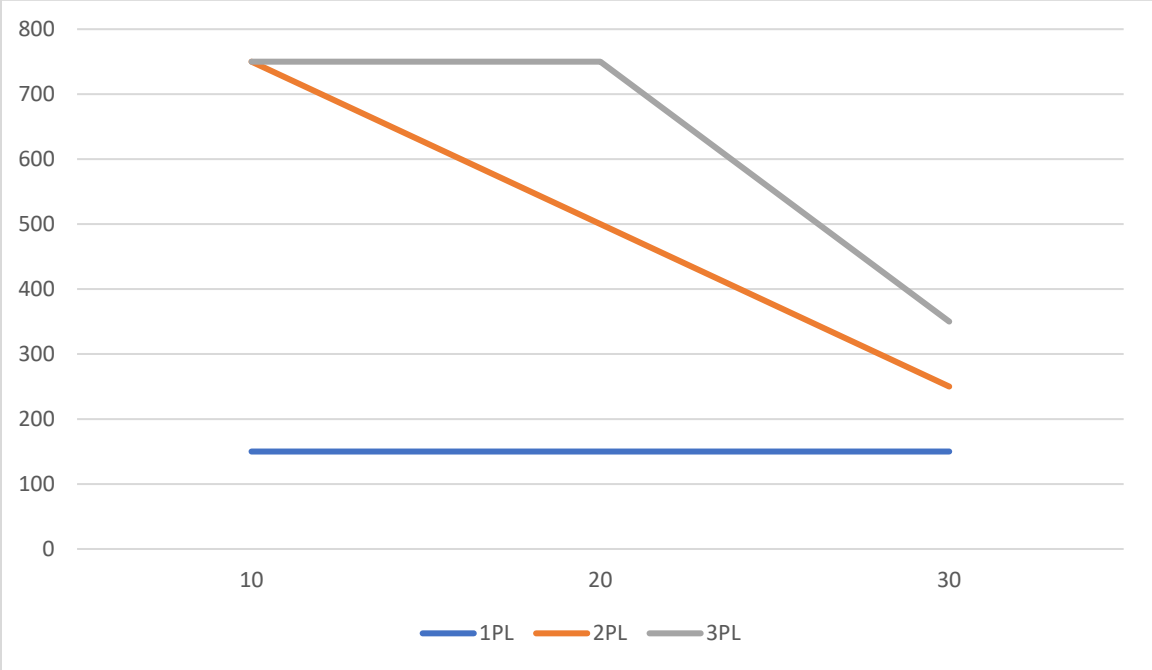
larger condition. Al-zboon et al. (2021) found that the amount of missing data has a significant impact on standard error, with more than 5% missing data presenting negative impacts. Han (2012) discussed theoretical particulars about interpreting guessing—harkening to the original caution against interpreting the third parameter strictly as guessing—and provided an empirical argument for fixing the c parameter in 3PL estimation to the probability of a random correct guess.

Sample Size

A decisionmaker presented with IRT as a potential information source might naturally ask the pragmatic question, "How many tests are required?" The answer—the complex, nuanced, and multi-faceted answer—might best be simplified into the less-than-actionable aphorism of, "It depends." Sahin and Anil (2017) aptly observe "tremendous discrepancies" in the literature, finding that research spanning 33 years offers evidence for anything from N=200 to N=1000 (p. 322). The latter number is an acknowledged rule-of-thumb, but the number of responses is a product of individuals and items. With subsets of a real test of English language proficiency prepared using exploratory factor analysis of items and random selection of individuals, under Marginal Maximum Likelihood Estimation, Sahin and Anil (2017) proceeded to evaluate product-moment correlations, root mean square differences, and χ^2 analysis to demonstrate minimum sample sizes of N=150 regardless of test length for the 1PL Model and minimum Item-Response products of around 8,333 for the 2PL and around 33,000 for the 3PL Model. Figure 3 summarizes these recommendations. Later, Uyar and Ozturk Gubes (2020) did not find significant problems when estimating a 2PL with 500 responses and 30 items based in part on those recommendations.

Figure 3

Recommended Minimum Sample Size by Test Length from Sahin and Anil (2017)



IRT models vary in complexity and cost. Table 1 summarizes some of them:

Table 1

Recommended Sample Sizes for IRT Models

Model	Introduction	Minimum Recommendations	
		Sample	Source
1PL	Rasch (1960)	150	Sahin and Anil (2017)
2PL	Lord et al. (1968)	500	Uyar and Ozturk Gubes (2020)
3PL	Lord (1984)	250-750	Sahin and Anil (2017)
GRM	Samejima (1968)	500	Reise and Yu (1990)

Estimation Procedures

The beneficial separation of person and item variables in IRT has an impact on how the specific numbers in models are determined; an established model may already have known item parameters, a new item's parameters might be estimated with responses from pre-measured

individuals, or a model may simultaneously need to estimate both item and person parameters. Hambleton et al. (2010) explain that, because of the assumption of local independence, the likelihood L of an observed response set u_n based on the underlying ability θ is the cross-product:

$$L(u_n|\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \tag{9}$$

and for mathematical reasons, a logarithmic transformation makes both the small values of L and ensuing operations easier to work with:

$$\begin{gathered} ln\, xy = ln\, x + ln\, y\,, \\ ln\, x^a = a\, ln\, x\,, \\ \therefore\, ln\, L(u|\theta) = \sum_{j=1}^{n} [u_j\, ln\, P_j + \left(1 - u_j\right) ln(1 - P_j)] \end{gathered} \tag{10}$$

and solving the first derivative of the resulting function returns the values for which likelihood is greatest. This equation cannot be solved directly even in the 1PL, though, because there are too many unknowns. Instead, various procedures can be applied to evaluate what is known fittingly as the "maximum likelihood estimate" (p. 35).

Advantages of Maximum Likelihood Estimation. There are pragmatic advantages to using maximum likelihood estimates beyond computational simplicity. When fitting a model for θ, for instance, the maximum likelihood estimator is normally distributed with a mean of the true value of θ and a variance expressed as:

$$V\left(\hat{\theta}|\theta\right) = [I(\theta)]^{-1} = \frac{1}{-E\left[\frac{\partial^2 ln\, L}{\partial \theta^2}\right]} = \sum_{i=1}^{n} \frac{P_i'^2}{P_i Q_i} \tag{11}$$

or the inverse of the information function I, where E is an expected value, and substituting the expected value of θ for the true value in the information function provides a confidence interval for the maximum likelihood estimator (p. 89). Hambleton and Swaminathan (1984) use the term "maximum likelihood confidence interval estimator" with a direct citation to Lord et al. (1968, p. 457) that, itself, further discusses the asymptotic efficiency of this estimator based on earlier

works (Cramér, 1946, p. 500; Kendall, 1961; Wald, 1942). Even Wald (1942) makes specific remarks to related, earlier research in Wilks (1938). In other words, the matter is a subject of long-standing research, and IRT models are *estimated* rather than *calculated*.

Algorithmic Estimation Procedures. Some of those estimation procedures, and the research presenting them, are well-organized and described in the work of Hambleton and Swaminathan (1984). These include the Newton-Raphson method, elaborated in Isaacson and Keller (1966), used to approximate the solution of the derivative of the log-likelihood such as when estimating person parameters given known item parameters:

$$\begin{gathered} let\ \frac{d}{d\theta} ln\, L(u|\theta) = 0 = f(x) \\ if\ f(x_0) \approx 0\ \ then\ |f(x_1)| < |f(x_0)|\ where\ f(x_1) = x_0 - h\ and\ h = \frac{f(x_0)}{tan\,\alpha} \\ tan\,\alpha = f'(x_0)\ \ \therefore\ \theta_{m+1} \equiv \theta_m - \left[\frac{d}{d\theta} ln\, L(u|\theta)\right]_m / \left[\frac{d^2}{d\theta^2} ln\, L(u|\theta)\right]_m \end{gathered} \tag{12}$$

where the slope of the likelihood function is used in an application of the distance formula and iterated until the improvement passes below a set threshold or the estimate "converges" (Hambleton & Swaminathan, 1984, p. 81).

In the 1PL in particular, starting values for θ (or in terms the above formula, x_0) can be established by taking the natural logarithm of the proportion of correct responses, and—more generally—perfectly correct and entirely incorrect response sets lack global maxima in their likelihood functions and are omitted from Newton-Raphson estimation with impacts that were then "currently not known" (p. 86). More specifically, the risk of small models yielding likelihood functions with multiple local maxima is dealt with by noting Lord (1980, p. 51) to support the claim that the risk is small "when working with large number of items ($n > 20$) as is usually the case in practice" (p. 88).

Joint Maximum Likelihood Estimation. The above Newton-Raphson procedure still

does not make the multivariate and nonlinear IRT models for unknown item and ability parameters estimable, because estimating item and ability (also called structural and incidental) parameters simultaneously creates a multivariate situation:

$$ln\, L(u|\theta, b, a, c) \;=\; \sum_{a=1}^{n} \sum_{i=1}^{n} [u_{ia}\, ln\, P_{ia} + (1 - u_{ia})\, ln\, Q_{ia}] \tag{13}$$

where θ, b, a, and c are now vectors of item parameters and u is a vector of responses with dimensionality Nn for N individuals on n items, and the Newton-Raphson procedure consequentially takes on a multivariate format to find the maximum via derivatives nudged to convergence in the similar implementation of the difference formula:

$$\begin{gathered} let\; f(t) = \; ln\, L(u|\theta, b, a, c) \\ let\; \frac{\partial\, ln\, L}{\partial t_k} = 0 = \; f'(t), where\; k = \; \{1 \ldots N + 3n - 2\}\; and\; t' = [\theta'\; b'\; a'\; c'] \\ if\; f'\left(t^j\right) \approx 0\; then\; \left|f'\left(t^{j+1}\right)\right| < \left|f'\left(t^j\right)\right| where\; f'\left(t^{j+1}\right) = t^j - \delta^j\; and\; \delta^j = \frac{f'(t^j)}{f''(t^j)} \end{gathered} \tag{14}$$

where t is a p-dimensional vector of parameters—constrained to length k to make it determinable—with the first derivative of the likelihood function producing a px1 vector and the second derivative providing a pxp matrix; the situation becomes mired in no fewer than eleven separate permutations of first and second partial derivatives for the 3PL with the possible introduction of Lagrange multipliers into the bargain. Hambleton and Swaminathan (1984) portray the situation aptly by noting that Lagrange multipliers are “rather complicated” without another word on the subject, settling on recommending a continued use of the two-stage iterative process of fixing certain parameters, converging, then fixing other parameters with the resulting convergence and repeating until changes are negligible (p. 132). Mathematical challenges to that process include local maxima, theoretically impractical convergence, and the Newton-Raphson procedure failing if the second derivative matrix becomes indefinite: using the information matrix instead of the second derivative is named as “Fisher’s method of scoring” (Hambleton & Swaminathan, 1984, p. 135; Rao, 1965, p. 302). The second derivative in this instance is a matrix

of partial derivatives called the “Hessian” matrix; researchers encountering this breakdown of Newton-Raphson convergence may recognize the name from the error message more than from the multivariate calculus.

Marginal Maximum Likelihood Estimation . Ultimately, the lack of sufficient statistics—that is to say, statistics that allow for the independent replication of mathematically equivalent samples—for most IRT models means that the two-stage iterative procedure of Joint Maximum Likelihood Estimation is theoretically inappropriate. Rather than finding the maximum of the likelihood function by solving the derivative to find a critical point, an integral can be constructed with respect to θ to eliminate the likelihood function’s dependence on its values. Darrell Bock and Lieberman (1970) is cited as one of the originators of this method (Hambleton & Swaminathan, 1984, p. 140). In their work, Darrell Bock and Lieberman (1970) decidedly state that such integration “overcomes the difficulty … posed by non-positive-definite tetrachoric correlation [or Hessian] matrices, which strictly speaking are not suitable for any form of common factor analysis” (p. 180) and describe the unconditional probability:

$$P(k = k_i) = \int_{-\infty}^{\infty} [\prod_{i=1}^{n} \Xi_{iki}(\theta)] \phi(\theta) g\theta \tag{15}$$

where Ξ is the likelihood of the response, not unlike $P_j^{u_j} Q_j^{1-u_j}$, φ is a normal distribution, and g appears to be an old formulation of d specifying the integral is with respect to—or perhaps given—θ. This is supported by the same formula being later described in Bock and Aitkin (1981, p. 445) as:

$$\begin{gathered} P(x = x_i|\theta) = \sum_j^n \left[\Phi_j(\theta_i)\right]^{x_{ij}} \left[1 - \Phi_j(\theta_i)\right]^{1-x_{ij}} \\ P(x = x_i) = \int_{-\infty}^{\infty} P(x = x_i|\theta) g(\theta) d\theta \end{gathered} \tag{16}$$

with similar variables while explaining how the Gauss-Hermite quadrature approximation as described by Stroud and Secrest (1966) allows approximating that integral as a sum of weighted,

indexed values that wind up corresponding to specific response patterns. They also suggest using the information function for the second derivative during the Newton-Raphson procedure—substantiating the claim with Rao (1965, p. 370) similarly to Hambleton and Swaminathan (1984)—but go on to their main point of explaining how the EM algorithm applies to and improves the process. Unfortunately, the explanation is dense: even the acronym of EM is not defined, but the importance of missing sufficient statistics for most IRT models is highlighted.

Generally, Dempster et al. (1977) originally defined the EM algorithm as Expectation Maximization, an iterated 2-step procedure suitable for estimating complete data given incomplete data; it can fit models like the 3PL when person and item parameters are both unknown, overcoming the challenge that insufficient statistics poses to Joint Maximum Likelihood Estimation. The EM algorithm is not unique to IRT; it is generally explained that:

$$g(y|\Phi) = \int_{\mathcal{X}(y)} f(x|\Phi)dx \tag{17}$$

where, to paraphrase in IRT terms, the observed data $\mathcal{Y}$ depend on latent parameters $\mathcal{X}$ and $g(y|\Phi)$ is maximized by leveraging the related $f(x|\Phi)$ to fit a complete model by integrating over sets of the sample space, estimate missing sufficient statistics, and repeating the process to refine the estimate of sufficient statistics until convergence is reached (p. 2).

More specifically, Bock and Aitkin (1981) continue to demonstrate how the EM algorithm can be used, further substituting conditional probabilities of θ for the maximization:

$$E(\theta|x_i) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1-\Phi_j(\theta)]^{1-x_{ij}} d\theta}{\int_{-\infty}^{\infty} g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1-\Phi_j(\theta)]^{1-x_{ij}} d\theta} \tag{18}$$

by using Bayes' theorem and approximating with response patterns as categories (p. 448). Muraki (1992) found the EM algorithm suitable to fit models even under a Generalized Partial Credit Model for polytomous responses.

Differential Item Functioning (DIF)

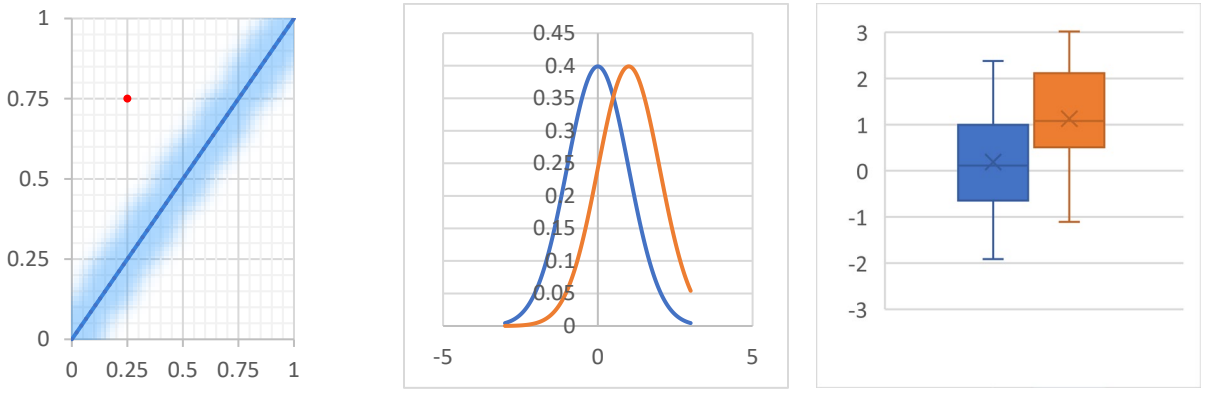
After sampling the models and estimation types present in IRT, a researcher might encounter several follow-up questions full of specific terms when attempting to empirically answer the question, "How good is this fitted IRT model?" The terms used are related both conceptually and mathematically, but they become more distinguished when considered carefully. Just as Regier et al. (2016) examined a large corpus of works from various climates to demonstrate that "local communicative needs" can explain the number of terms for snow that Franz Boas famously acclaimed in the Eskimo languages, so too might the various, related terms used to describe model quality arise from needs within the field. Thus, before examining DTF, the construct of Differential Item Functioning (DIF) is best described by briefly examining the nature of statistical difference and testing power.

Statistical Difference

It is helpful to establish the statistical algorithm of looking for significant differences when considering DIF testing. A rather unoriginal example of a statistical difference problem is the fishing question: assuming the lengths of fish from two fishing trips are available, can it be said that the trips were to different destinations? Any one of the following figures might be drawn during the discussion:

Figure 4

Depictions of statistical difference



While the various figures emphasize different concepts, from the assumption of hypotheses to the nature of the sample, the initial question could be reduced to arithmetic—to the division of differences—to provide a quantitative substantiation of some evaluative heuristic.

When evaluating the quality of an IRT model, different perspectives similarly draw on different facets of the same mathematical whole. A researcher concerned with statistical power might look at probabilities of differences. One concerned with reliability might look at a signal-to-noise ratio. Yet another concerned with model fit might be concerned with the parsimony of comparative error rates. Just as a statistical test requires its null and alternative hypotheses, so too do these measures of model quality require their local, even linguistic, context. Traditionally, DIF is considered a problem that can be measured with hypothesis testing: the question is usually, “Given these groups and their measurements, can the null hypothesis that there is no difference between them when it comes to this IRT model?”

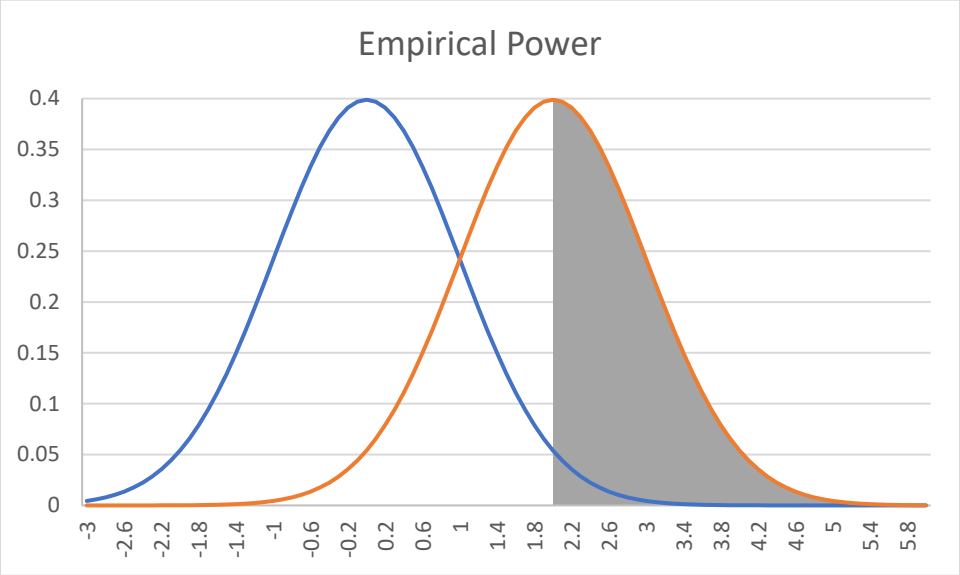
Testing Power

The traditional hypothesis testing for DIF statistical power; the process to reject the null hypothesis uses cut-off values to attain a desired power. Statistical power was well-described in Wright (2011, p. 73) as “the probability of correctly rejecting a false null condition” and, more

visually, empirical power as “the area beyond the null critical value, under the alternative distribution” (p. 52) as roughly depicted:

Figure 5

Empirical Power



where, technically, the shaded area under the null hypothesis curve indicates a Type I error. Since power—statistical and empirical—is concerned with hypotheses, it can be thought of as a trait of statistical tests. Not every statistic produces a statistical test, and not every statistical test has a known power. Conversely, statistical tests with well understood power can support a priori calculations to determine minimum sample sizes required for statistical tests. Software such as G*Power helps automate this process (Faul et al., 2009).

Appropriate Dimensionality. Another way of examining DIF is to determine if it effectively measures only intended—usually one—dimension(s). Explicit evaluation of the unidimensional measurement of an IRT model is often concerned with interitem correlation involving the tetrachoric matrix (Lord & Novick, 1968; McBride & Weiss, 1974, p. 30; Penny, 1994, p. 31). Geometry plays a role here: the basic formula for r_{tet} makes use of a cosine:

$$r_{tet} = cos\frac{\pi}{1+\sqrt{OR}} \quad (19)$$

where OR is the odds ratio, and the cosine is in radians. Generally, the tetrachoric matrix can be observed in the construction of odds ratios used in testing the suitability of models.

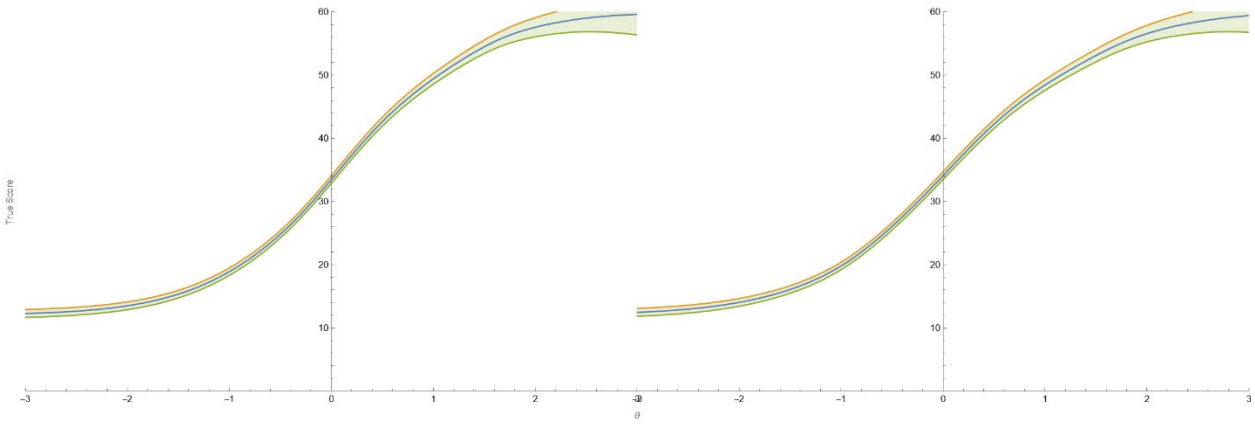
Standard Error of the Estimate. Item Response Theory also has traditional measurements of Standard Error. Culligan (2011) elegantly traces the evolution standard error through Classical Test Theory, reliability measurement, logits, and the 1PL before observing:

$$SE(\theta) = \frac{1}{\sqrt{\sum_{i=1}^{n} P_i(\theta)Q_i(\theta)}}$$

where the denominator is the "Item Information Function" (p. 7). The challenge of IRT's non-constant Standard Errors can be visually appreciated by applying the formula to the instrument used in Duncan (2006) to produce Test Characteristic Curves:

Figure 6

Test Characteristic Curves with Emphasized Standard Errors



The formula also has a particular virtue when noting that the variance-covariance matrix at model fitting convergence provided by software such as R's *ltm* package can be used as well. This measurement, however, lacks the declarative ability of a test-based statistic.

Differential Item Functioning Statistics

If an item does not effectively measure what it is intended to, then it is unreliable. If some of the unintended measurement comes from some other property of the subject rather than to random error, then the item—and indeed, perhaps even the test it is part of—functions differently for different individuals, violating a core assumption of IRT and potentially creating problems for the fairness of tests. That item would then exhibit what is called Differential Item Functioning. Many approaches exist to test for these sorts of problems. Draxler (2010, p. 708) synthesizes the efforts of C.A.W Glas and N.D. Verhelst (1995) to enumerate broad categories of test types used for Rasch models as "χ^2 tests, likelihood ratio tests, Wald tests, and Lagrange multiplier tests" while also observing that all these sorts of tests do not control for Type II errors; they are concerned with hypotheses that involve items performing significantly different for subgroups of participants and consequentially prioritize statistical power and minimization of Type I error. The hypotheses necessary for these tests tend to be formed based on a subgroup's different model performing better than those of the population at large.

Chi-square Tests. The χ^2 distribution is, in fact, more than one distribution; given a normally distributed variable, what should the values of k samples look like? Kissell and Poserina (2017, p. 120) efficiently characterize the χ^2 distribution as "the distribution of a sum of squared random variables" where k, the degrees of freedom, is equal to the number of random variables in the theoretical sample distribution. These properties allow for calculating expected values, and, in turn, the observed values can be assessed for their probability. Multiple evaluations are made using this distribution throughout Item Response Theory. Liang and Wells (2009) observed that measures of model fit using the χ^2 statistic including over-sensitivity with large samples – which interacts poorly with the large samples require to fit more complex

models in the first place – inflated Type I error rates when grouping examinees on θ, and potential misspecification of degrees of freedom for χ^2 under the null hypothesis. The latter two points are substantiated with the work of Orlando and Thissen (2000) and continue to extend the criticisms even to log-likelihood procedures using χ^2 in the later works of Orlando and Thissen (2003) and the Lagrange multiplier test examine by Glas and Falcón (2003). Glas and Falcón (2003) additionally noted that missing data can make Type I error rates "undesirably high, especially for short tests" in these tests (p. 87).

Mantel-Haenszel Method. A χ^2 with one degree of freedom can be conducted on the "difference between the cases [A and B] and controls [C and D] in the proportion of individuals having the factor under test [A and C]," such that:

$$\frac{\left(|AD-BC|-{}^{1}/_{2}T\right)^{2}T}{N_1M_1N_2M_2} \tag{20}$$

or, somewhat more accessibly, as testing the assertion:

	With factor	Free of factor
With disease	P_1	P_2
Free of disease	P_3	P_4

$$\frac{P_1}{P_2} = \frac{P_3}{P_4} \tag{21}$$

where the population is divided into four distinct proportions (Mantel & Haenszel, 1959, p. 730). Holland and Thayer (1988, pp. 130-135) combined this with other contemporary work to apply the procedure to IRT models:

Group	Score on Studied Item		Total
	1	0	
R	A_j or p_{Rj}	B_j or p_{Fj}	n_{Rj} or 1
F	C_j or q_{Rj}	D_j or q_{Fj}	n_{Fj} or 1
Total	m_{1j}	m_{0k}	T_j

$$MH - CHISQ = \frac{\left(|\Sigma_j A_j - \Sigma_j E(A_j)| - {}^{1}/_{2}\right)^{2}}{\Sigma_j Var(A_j)} \tag{22}$$

where each item is examined for DIF, but a common factor among all items can be determined:

$$\hat{\alpha}_{MH} = \frac{\sum A_j D_J / T_j}{\sum B_j C_J / T_j}, and\ \Delta_{MH} = -\frac{4}{1.7} ln(\hat{\alpha}_{MH}) \quad (23)$$

to create a scaled measurement of DIF centered around 0. In addition to evaluating DIF itself, the MH statistic has also been used to develop effect sizes for other measurements of DIF (Wright & Oshima, 2015). This method has been found to outperform Lord's χ^2 and likelihood ratio tests (Diaz et al., 2021). However, sample sizes remain important: Mazor and et al. (1991, p. 8) found over half of DIF items (mostly, difficult or poorly-discriminating items) were missed in some conditions and "there would seem to be little justification for using sample sizes any smaller than 200" with the MH procedure as a result. MH has been found to be effective even in polytomous IRT models (Wen-Chung & Ya-Hui, 2004). Wright and Oshima (2015) developed a measure of effect size for Differential Functioning of Item and Test analysis. While discussing similarities in the performance of NCDIF and MH statistics, the research observes that Mantel-Haenszel techniques overestimated DIF for easy and difficult items due to area-based measures failing to account for sample sizes present in areas of the ICC curve; in other words, it appears that cell loadings may be a concern.

Likelihood Tests. Wilks (1944, pp. 150-152) described a ratio where:

$$\lambda = \frac{P_\omega(O_n)}{P_\Omega(O_n)} \quad (24)$$

such that O_n is a particular parameter, and λ is the likelihood ratio of the null hypothesis that P_ω and P_Ω are equivalent; it is shown that "the likelihood ratio test for H_0 is seen to be the same as the (Student) [sic] test" and $-2 \log \lambda$ follows a χ^2 distribution with degrees of freedom equal to the number of probabilities assumed to be equal under the null hypothesis (e.g. 2 for a reference and single focal group.) Waller (1981, p. 119) propose using the likelihood ratio test for a "goodness-of-fit statistic" to compare IRT models. Brown et al. (2015) criticize likelihood ratio tests between the 2- and 3-parameter IRT models because setting a guessing parameter to zero to

compare the 2PL as a special 3PL "violates one of the assumptions of the likelihood ratio test and renders the usual χ^2 distribution inappropriate for the comparison" (p. 335) and empirically demonstrated inaccurate Type I error rates; they suggest both stepwise removal of individual item guessing parameters and using p-values from a null distribution rather than a standard reference distribution for likelihood ratio tests. Both χ^2 and log-likelihood tests of contingency tables—or tetrachoric matrices—are problematic when cells—or expected values—are small (Bartholomew & Shing On, 2002; Cai, 2008, p. 314).

Wald Tests. The Wald test, explained in Wald (1943) with mathematical proofs, has been applied to assessing model fit in Item Response Theory (Draxler, 2010; Fischer, 1995). In particular, the Wald test informed development of the Lord (1980) χ^2 test. IRTPRO software makes use of the combination of Lord's χ^2, concurrent calibration as detailed in Kim and Cohen (1998), and the supplemental expectation maximization algorithm from Cai (2008) to evaluate DIF without selecting anchor items in the Wald-2 test option. Wald-2 provides *concurrent calibration* where reference group means are fixed normally to estimate focal group parameters, then the focal group parameters are fixed to that estimate in a looping procedure similar to the EM algorithm with advantageous performance noted in unequal sample sizes (Langer, 2008; Woods et al., 2013). Concurrent calibration procedures benefit from iterative estimation or other purification procedures before the final analysis is conducted (Fikis & Oshima, 2017; González-Betanzos & Abad, 2012; Kopf et al., 2014a, 2014b; Wang, 2004; Woods, 2009).

Area Tests. Lord (1980) pointed out that when there is no DIF, the Item Characteristic Curves for reference and focal groups are congruent with the overall ICC. Afterwards, Raju (1988) developed and proved formulae for computing the area between two item characteristic curves for a focal and reference group on a given item:

$$Signed\ Area\ (SA) = (1-c)(b_2 - b_1)$$
$$Unsigned\ Area\ (UA) = (1-c)\left|\frac{2(a_2-a_1)}{Da_1a_2} ln\left(1 + exp\left(\frac{Da_1a_2(b_2-b_1)}{a_2-a_1}\right)\right) - (b_2 - b_1)\right| \quad (25)$$

where the curves are on the same scale and have identical guessing parameters (p. 496). These proofs informed the development of the Differential Functioning of Items and Tests approach established in Raju et al. (1995) which also includes examining Test Characteristic Curves for Differential Test Functioning. Penny (1994) used the Mantel-Haenszel χ^2 statistic to evaluate Differential Item Functioning (DIF) in the difference of definite integrals of Item Characteristic Curves. A 3-Paramter, unidimensional, monotonal model was fitted to data using maximum likelihood estimators in LOGIST. Although the tests were found feasible, the study observes that the method has different amounts of sensitivity for different parameters and raises the question of using the Mahalanobis distance as a response. The study also notes that demonstrably aberrant items are worth omitting in DIF Analysis; the error these items add to measurements can obfuscate DIF in other items.

Anchor Item Selection. If IRT models are tested for DIF based on separately estimated parameters, such as with many procedures that compare models to test a null hypothesis, then it is necessary to ensure that those parameters are on the same scale. This is done traditionally by "anchoring" reference and focal group estimates to the same scale by constraining items in common to both datasets, in essence using particular item(s) for calibration of θ. During some of the procedures developed to test for differential item functioning, however, the situation becomes theoretically complex. When using a test statistic to identify anchor items for subsequent tests, Type I and Type II errors can change roles when these tests are used to screen intermediately for proper items rather than identity improper ones (Fikis & Oshima, 2017). Schulze et al. (2022) suggest averaging model parameters using Bayesian methods as a test of measurement

invariance when anchor items cannot be specified and comparing between groups. Robitzsch and Ludtke (2022, p. 59) further examine linking approaches with PISA data and add to the body of evidence that approaches which do not force invariance assumptions to scale estimates provide superior DIF detection; they even observe computational times of 10 minutes reduced to "at most 3s" with such algorithms.

Comparative Studies. Writing in the context of quality-of-life healthcare measurement, Jin-Shei et al. (2005) identified three methods for DIF analysis in small sample size environments favoring the 2PL. The first was a 1-PL IRT/Rasch method based on t-tests between item parameters with logit differences where N>200, citing the work of Wright and Panchapakesan (1969), and they "adapted the concept of the MH D-DIF index classification system" to set logit thresholds for borderline DIF (Jin-Shei et al., 2005, p. 288). The second was logistic regression, citing Crane et al. (2004) and Zumbo (1999), with significant χ^2 differences in models with and without focal group identifying DIF to interpret with a 0.13 ΔR^2 cut-off as an effect size; they observed that effect size remained an opportunity for further research. Finally, they mentioned the DFIT framework, citing Flowers et al. (1999) and Raju et al. (1995), noting the practical appeal of analyzing differential function at the test level and suggest χ^2 testing to evaluate the significance of any differences between focal and reference group functioning; DFIT can be interpreted as a reliability measure because it is formed from non-compensatory DIF. Jin-Shei et al. (2005) note that within their line of research, using more than one DIF analysis method is "common" (p. 291).

Fikis and Oshima (2017) investigated an algorithm for identifying and omitting aberrant items prior to conducting DIF Analysis. A 3-Parameter, unidimensional, monotonal model—albeit with a fixed guessing parameter—was fitted to data using marginal maximum likelihood

estimators in IRTPRO. The software ships with the Wald-2 anchorless DIF test, which combines samples and instruments by recoding unshared items as missing data in a single model. The use of the algorithm resulted in more effective DIF Analysis, but the increased calculations added undesirable computational complexity. The study questions the theoretical implications of using statistic tests in ways they were not originally designed to be used; the theoretical roles of Type I and Type II error may alter when using a test for identification as a test for omission.

Elosua and Wells (2013) examined, among other things, the performance of mean-covariance structure, ordinal logistic regression, and likelihood ratio tests to detect DIF in a polytomous IRT model. They found that likelihood ratio tests exhibited increased Type I error rates and theorized that items "may be corrupted," (p. 339) suggest examining how anchor item selection impacts the performance of the test, and recommend the generalized Mantel-Haenszel method of Fidalgo and Madeira (2008) and Fidalgo and Scalon (2009) for non-uniform DIF.

Ippel and Magis (2020) developed improved asymptotic standard error formulae for weighted likelihood and Bayes modal estimation of θ which outperformed existing methods, and it was additionally found that Exact Standard Error exhibited superior performance in all conditions; Exact Standard Error is:

$$SE_{exact}(\hat{\theta}) = \sqrt{\sum_{t=1}^{T} p^{(t)}(\hat{\theta})\left(\hat{\theta}^{(t)} - \bar{\theta}\right)^2} \tag{26}$$

where t represents a single response per each possible response set—with 2^n possibilities for an n-item test—that create a sample for estimating $\hat{\theta}$ and "$\bar{\theta} = \sum_{t=1}^{T} p^{(t)}(\hat{\theta})\hat{\theta}^t$ is the average ability estimate of the sample distribution," but standard error for item parameters was unexamined (p. 466).

Differential Test Functioning (DTF)

The work of Raju et al. (1995) brings together the area-based approaches to reliability measurement and DIF testing with a framework that considers impact at the level of the test; since decisions are made based on test scores, it naturally follows that measurement discrepancies are most problematic when they affect the overall instrument. The concept of compensatory differential item function (CDIF) is examined in traditional IRT models, polytomous, and multidimensional contexts by existing literature (Flowers et al., 1999; Oshima & Morris, 2008; Raju, 1988; Raju et al., 1995).

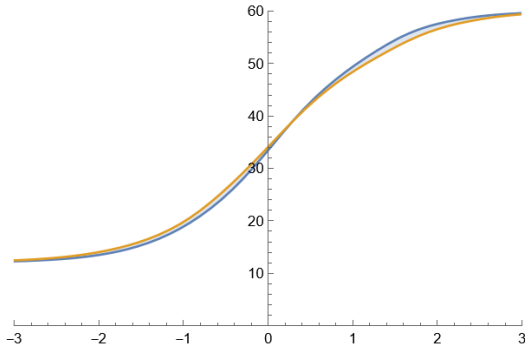
The DFIT framework includes multiple formulae and algorithms, including:

$$DTF = E_f[D(\theta_s)^2] \tag{27}$$

where "the expected value of the squared difference between focal and reference groups, where the expectation is taken across the θ distribution from the focal group," which is more easily understood when depicted visually for the Test Characteristic Curves for the instrument in Duncan (2006) data:

Figure 7

DTF depicted as the area between reference and focal TCFs



and, consequentially, the guessing parameter must be the same in both models to prevent the area between curves from being infinite.

DIF, and DTF, have enjoyed continued development much like the rest of IRT; in the course of examining a DIF detection method using explanatory covariates, De Boeck and Cho (2021, p. 712) describe the state of the literature as an "unstoppable continuation of DIF method research." Significance testing under the DFIT model originally followed an iterative procedure where items with CDIF were pruned from the model until DTF is found to be insignificant via a χ^2 testing; the non-retained items become flagged as DIF items (Oshima & Morris, 2008; Raju et al., 1995). Later developments included the Item Parameter Replication (IPR) procedure where cut-off values for NCDIF were empirically determined for a particular test scenario by simulating multiple sets of estimated parameters with constrained variance-covariance matrices to generate a null hypothesis distribution that can provide cut-off values for desired power levels (Oshima & Morris, 2008; Oshima et al., 2006). Further work has developed effect size measurements for the NCDIF statistic; not only can a significant difference be empirically determined, but it can also be practically interpreted (Wright, 2011; Wright & Oshima, 2015). It has been noted that Item Parameter Replication benefits from using variance-covariance matrices for both the reference and the focal group as opposed to just the reference group (Cervantes, 2012; Cervantes, 2017). The variability difference problem has also been addressed by using a weighted distribution to improve DTF, with observable improvements in the Type I error rates as sample size increases (Chalmers et al., 2016). The overall performance of IPR critical values has also been empirically verified, with results observing lower power when DIF was because of discrimination (the *a* term in the 2PL) differences and overall benefits from iterative linking (Seybert & Stark, 2012). Similar approaches to IPR's method of generating data under a null

hypothesis for comparison in a D-QQ plot has also found potential power gains while avoiding the necessity of anchor items (Yuan et al., 2021).

While advantageous and methodical, using DFIT with IPR cut-offs to answer the question, “Does this instrument have problems?” requires a small simulation study *for each iteration*. Each time DTF is measured to check for insignificance, the cut-off values must needs be redetermined with fresh variance-covariance matrices from the smaller and smaller set of items as determined by their CDIF. This pruning method carries the virtues of purification methods as examined in Fikis and Oshima (2017) but necessitates even more computational complexity and care with anchor item selection. Metaphorically, it is like carving a pair of shoes from wood: the shoe must be tried on by more than one wearer and problematic bits carved away, then retried by another group of wearers, and potentially worked again until the fit is right for use by a pre-determined proportion of wearers, with whatever is left on the floor deemed as undesirable in the final product. The fit might be great, but so is the cost.

Mahalanobis Distance

In some ways, it is tempting to perceive DIF as a reliability problem, because in some ways, that is correct. Unfortunately, though, most of the methods to find DIF are statistical tests that operate under a hypothesis test framework requiring specifications and data beyond the original scope of their original models. The conditions that may warrant a DIF analysis are more poorly defined and investigated than the conditions required to conduct a DIF analysis. Even when these issues are dealt with, the threat of bad anchors has created a need for additional computation through the process of item purification. Item purification itself has theoretical issues to consider based on tests of null hypotheses being used to ultimately retain, rather than reject, items. Procedures such as Item Parameter Replication require even more computation to

iterate through cut-off values. The researcher looking over various formulae would recognize the common themes of significant statistical difference and the presence of error in estimation, and one might be tempted to solve these two and more challenges by combining them.

The Mahalanobis distance determines multivariate normality by reducing even potentially correlated variables to a one-dimensional or *Euclidean* distance; distances are transformed by measuring them from the multivariate point of a particular centroid in the same space after accounting for their correlation (Mahalanobis, 1936). Outliers are datapoints which are outside "a swarm around the centroid in multivariate space" formed by most of the data (Fidell & Tabachnick, 2003, p. §3.2.3). Rao (1965) describes the test of Mahalanobis distance as:

$$D_p^2 = \sum \sum s^{ij} d_i d_j \tag{28}$$

where d is the difference in sample means of multiple variables of different populations (p. 480).

The Mahalanobis distance is well-suited to providing a reliability statistic capable of flagging IRT models for more thorough DIF analysis. Its potentially simple application in χ^2 testing could provide the significance testing that DFIT solved only through the extensive Item Parameter Replication purification framework developed in Oshima et al. (2006). Indeed, item purification for DIF detection has even received some critical scholarship, such as the supposition by Magis and Facon (2013, p. 309) that "item purification does not improve both the Type I error control and the power" after showing how Delta plots work as well or better in smaller sample sizes. Its multivariate scaling properties can help account for the differences in IRT parameter performance that can lead to challenges in DTF calculation such as the lower power in cases of DIF based on the discrimination (or *a* parameter in the 2PL) observed in Seybert and Stark (2012). Furthermore, it can potentially be constructed to avoid the significant challenge of both anchor item selection described by Dimitrov (2017) and measurement

invariance between reference and focal groups as enumerated by Chalmers et al. (2016). Penny (1994, p. 100) suggests further research using the Mahalanobis distance, observing that its unweighted comparison of item parameters might aid in "providing a direct assessment of DIF as defined by Lord (1980)." That decorated work of Lord (1980)—if one might call over 2,000 known citations a decoration—includes specific mention of using the Mahalanobis distance to directly compare models' parameters. Unfortunately, though, none of those works citing Lord (1980) mention the Mahalanobis distance in an accessible manner. The approach remains less-investigated, much like the 4PL mentioned in Barton and Lord (1981).

Bechger and Maris (2015) applied the statistic to the detection of DIF. In their study, they propose that comparing items' relative parameters is a more meaningful way to look for DIF than individual items, calling it "differential item pair functioning." The case for this approach is made when considering a few traits of models. First, the way that IRT models produce estimates with varying scales for parameters requiring anchoring between reference and focal groups for meaningful comparisons. Second, existing DIF methods are known to be sensitive to true group differences and they tend to display inflated Type I error rates by flagging them as DIF items. Overall, they make a point of noting how altering the composition of an instrument can change its presentation of DIF. Their proposal is based on this theory: because the statistics that would be required for a traditional null hypothesis are not what is estimated in reference and focal group models, they re-arrange the null hypothesis as a zero difference in relative difficulties rather than as an equality between difficulties.

Their solution is to create a matrix of pairwise differences of parameters between groups. After pointing out some met assumptions on multivariate normality and independence, they

observe that testing the matrix they created is accomplishable by a χ^2 test of Mahalanobis distances:

$$\chi^2_{\Delta R} \equiv \hat{\beta}^T \sum^{-1} \hat{\beta} \quad (29)$$

where β is an arbitrary column of the difference between parameters of the two models with degrees of freedom 1 fewer than the number of items. They conclude by presenting an omnibus DIF test based on simultaneous pairwise comparisons of these differences, grouping them into "clusters," where DIF items form isolates. Importantly, they demonstrate with a real-world dataset that choosing bad anchor items mis-identifies DIF. The study notably indicated that samples of less than 500 in reference or focal groups presented challenges to DIF identification. Their "main message … is that DIF can only be defined in terms of the identified parameters" (337). Pohl et al. (2021, p. 489) extend this work by applying k-means cluster analysis to further classify invariant items' difference in relative item difficulties, offering what was called the "first comprehensive results" of such an application. The Mahalanobis distance is also frequently mentioned in multivariate research (Tabachnick & Fidell, 2007). Specific mention of the Mahalanobis distance in DIF research, though, is currently sparse.

In conclusion, the current study thus proposes using the Mahalanobis distance, which has been both suggested and effectively used in the literature, to examine standard errors, which themselves have seen some use in the literature in various forms, to assess whether an instrument's items should be tested for DIF.

3 METHODOLOGY

Overview

The purpose of this study was to compare the performance of a computationally advantageous statistic to the traditional DTF statistic. The proposed new statistic is the skewness of the Mahalanobis distances of the standard errors of an estimated model. It is theorized that it may function as an indicator of potential differential item functioning (DIF) that, unlike traditional DIF analysis methods, requires neither the theoretical framework of an alternative hypothesis nor the practical framework of larger samples and multiple models. Consequentially, this statistic might add value as a routine "check engine light" to indicate the need for more testing.

To conduct that comparison, a simulation was conducted to measure both statistics' values and computational costs in some typical IRT conditions. Correlations of the values were then compared, and the nature of any correlation across simulation conditions was analyzed.

Monte Carlo Simulation Method

The presumption of simulation as a valid method of inquiry presents risks of creating a *hyperreal* which may alter the relationship between observation and observer by displacing reality with a new, meaningful construct (Baudrillard, 1994). Even so, the use of Monte Carlo (MC) estimation and simulation studies in IRT is well-established. Harwell et al. (1996) summarized methods and concerns up until that point, in particular noting that, "perhaps the most popular outcome variable in MC studies in IRT is the root mean square deviation (RMSD)" (p. 270). Millsap (2011, p. 167) later observes that even Markov Chain Monte Carlo estimation approaches continue to "have grown in popularity" in real-world IRT applications as a means of dealing with its complex, multivariate distributions.

Using Standard Errors for the new Statistic

Lord and Novick (1968, p. 374) theorize that, among other things, a normally distributed, unidimensional "latent space" of θ will yield a response set U which is multivariate normal. This theory does not, however, suggest that every test is made of items with normally distributed parameters. Thus, while it may be initially tempting to consider parameter estimates as multivariate values that can be tested for multivariate normality, such an assumption is perilous. Even if the standard error of parameters is theoretically independent, the values in the Hessian matrix lack independence, technically making χ^2 tests inappropriate. The Standard Error remains theoretically suited to describing the accuracy of a parameter estimate. This study used the skewness of the standard errors rather than testing for their normality as a result.

Using DTF for Comparison

Differential Test Functioning (DTF) is part of the Differential Functioning of Items and Tests (DFIT) framework that can provide measures of both Compensatory Differential Item Functioning (CDIF) and Non-compensatory Differential Item Functioning (NCDIF) for items (Raju et al., 1995). It can be calculated by taking the area under the difference of test characteristic curves:

$$DTF = E_f\left(D^2_{i|W}\right) = \int D^2_{i|W}\, g_f(W) dW \tag{30}$$

and, "in practice, DTF is defined in relation to focal group members only" (Millsap, 2011, p. 224). This means that calculating DTF requires linking parameter estimates, potentially specifying anchor items, and identifying focal and reference groups. DTF is used within the DFIT framework as a pruning statistic: in the IPR method, procedures are conducted to eliminate bad items until DTF appears satisfactory (Oshima & Morris, 2008; Oshima et al., 2006).

Design

Instrumentation, sampling, and conditional components of the simulation were based on Fikis and Oshima (2017) which, in turn, was informed by in Seybert and Stark (2012) to evaluate conditions with various types of typical DIF. Conditional, replication, and analytical components of the study were further informed by related research. Conditions both of interest from the perspective of IRT modelling and specifically of DIF itself were included. This study also includes conditions related to the imperfect specification of models during DIF analysis.

Instruments: Test Length and DIF Amount Conditions

Simulation conditions replicated those of Fikis and Oshima (2017) and Seybert and Stark (2012) by using two forms of instruments, one a 30-item test and another a 15-item subtest of the same instrument using the following parameters in the 3PL. The 15-item subset was determined based on the subset used in the previous research:

Table 2

Item Parameters

Item	Difficulty (b)	Discrimination (a)	Pseudo-Random (c)
1	-0.07	0.49	0.19
2[a]	0.21	0.92	0.15
3	0.54	1.26	0.05
4[b]	-0.03	0.61	0.18
5	0.01	1.74	0.12
6	1.96	0.5	0.12
7[b]	0.04	0.96	0.13
8	-0.09	0.59	0.18
9	-1.16	0.82	0.17
10[a]	0.02	1.26	0.11
11	0.2	0.82	0.07
12	-0.43	0.75	0.15
13[b]	-0.06	1.49	0.09
14	-0.34	0.97	0.12
15	0.05	1.49	0.12
16	-0.25	0.89	0.15
17[a]	0.06	1.45	0.07
18	0.31	0.75	0.18
19[b]	0.04	1.43	0.08
20	0.13	0.6	0.22
21	0.52	0.83	0.09
22[b]	-0.96	0.56	0.19
23	-0.79	0.67	0.2
24	0.37	0.7	0.18
25[a]	-0.71	1.03	0.14
26	-0.19	0.89	0.21
27	0.74	1.23	0.06
28[b]	-0.44	0.9	0.18
29	-0.17	1.23	0.12
30	0.53	0.69	0.17

where a notes DIF items under all DIF conditions and b notes DIF items under 33% DIF; DIF is simulated by altering the difficulty parameter of the focal group.

Sampling: Sample Size and DIF Magnitude Conditions

Sample sizes were chosen both from theoretical and practical perspectives. The initial sample size conditions in Fikis and Oshima (2017) were both informed by the recommended

sample sizes for model estimation in Sahin and Anil (2017) and Uyar and Ozturk Gubes (2020) as well as by the typical American school's combined grade level size of 50 based on *Common Core of Data (CCD)* (2021).

Sampling also included a condition to vary the magnitude of DIF to generate observable DIF. Wright and Oshima (2015) observed that impact could affect the evaluation of DIF while discussing "comparing the comparables" (p. 7) as recommended by Dorans and Holland (1992). Simulation participants were generated from random normal distributions—$N(0,1)$—but the difficulty parameters for DIF items was manipulated between 0.5 and 1.

Other Properties

Generating Model Conditions. Cuhadar et al. (2021) identified some significant consequences to failing to appropriately specify the pseudo-random C-parameter in IRT models when it comes to measurement invariance analysis; consequentially, conditions including ignoring, fixing, and estimating this third parameter were included in the study.

Replications. Pekmezci and Avsar (2021) investigated the number of replications appropriate for simulations in unidimensional IRT models and empirically recommend at least 625 replications for studies concerned with Type I error rates for all models concerned. Because of this recommendation and given the simplicity of the pre-screening statistic under investigation, 625 replications were used.

Procedures

This study proposed a simulation of 120 condition permutations with 625 repetitions per condition. The study combined the percent of DIF and the magnitude of DIF into one condition to retain a fully factorial design structure:

Table 3

Simulation Conditions

Condition	Values	Inspiration
Test Length	15	Fikis and Oshima (2017)
	30	
Sample Size	50	*Common Core of Data (CCD)* (2021)
	250	Sahin and Anil (2017)
	500	Uyar and Ozturk Gubes (2020)
	1000	Fikis and Oshima (2017)
DIF Presence	(None)	Wright and Oshima (2015)
	20% @ -0.5	Fikis and Oshima (2017)
	33% @ -1	
	20% @ -0.5	
	33% @ -1	
Model Selection	2PL2	Cuhadar et al. (2021)
	PL with fixed c	
	3PL	

and the general procedure in each repetition was as follows:

1. Generate random item responses based on test length, sample size, DIF percentage, and magnitude condition using a model based on the model condition where for a 3PL:

$$P(\theta)_{ij} = \sum_{j=1}^{n_j} \sum_{i=1}^{n_i} P_i(\theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}$$

   or C = .2 for the 2PLC condition, or C = 0 for the 2PL condition, then:

   a. Begin with θ_j, a random normal number per participant
   b. D, the scaling parameter, is 1.7
   c. Prepare item parameters b, a, and c for the reference group by following the specifications in Table 2 and the number of items condition.
   d. Prepare item parameters b, a, and c for the focal group the same way, but additionally increase the b-parameter for items identified with DIF

according to the DIF condition (0%, 20%, 33%) by an amount from the magnitude condition (0.5, 1.)

e. Randomly select 25% of the sample to classify as the focal group.

f. Calculate $P(\theta_{ij})$ for each item response

g. Compare $P(\theta_{ij})$ to a random uniform number; if $P(\theta_{ij})$ is greater than the random uniform number, then the item response is coded as 1 for a correct response. Otherwise, it is coded as 0 for an incorrect response.

The result is one matrix per 625 repetitions per 120 conditions of width equal to the number of items (15 or 30) and length equal to the sample size (50, 250, 500, 1000) condition values, or 90,000 matrices of dimensionality anywhere from 15x50 to 30x1000. Other conditions are applied during later analysis stages.

2. Estimate a full model based on a 2PL:

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$$

The ltm package provides methods for estimating these models. Each repetition will include a full model using all items and all participants in the simulated data; this represents the "default settings" option an everyday practitioner might use.

3. Determine the skew of Mahalanobis distance of standard errors for the full model:

a. Evaluating the Hessian matrix $H(\hat{\theta})$ at model convergence.

b. Inverting $H(\hat{\theta}_{ML})$ and find square roots of its trace to approximate the standard error of estimated parameters, $SE(\hat{\theta}_{ML})$

c. Separating $SE(\hat{\theta}_{ML})$ by item and parameter for a multivariate dataset, for example a 2PL fit to 30 items will have 30 points of 2-dimensional data

   d. Determining the Mahalanobis distances for each item's $SE(\hat{\theta}_{ML})$ by:
      i. Determining the centroid by averaging on all variables, and
      ii. Measuring the distance of each item's Standard Errors from it
   e. Noting the skewness of the Mahalanobis distances for each model

   The ltm R package includes a method for extracting the standard errors under a traditional IRT parameterization, the e1071 package provides a skewness function, and the base stats package includes a Mahalanobis distance function.

4. Estimate DFIT models (equated focal and reference) using the *ltm* R package.
   a. Specify reference and focal groups.
   b. Fit an IRT model for the reference group's responses.
   c. Choose three DIF-free items at random as anchors.
   d. Constrain focal group item parameters for anchor items to values found for the reference group to scale θ values.
   e. Determine DTF using the *dfit* R package for model parameters fit with *ltm*
5. Examine the time taken to conduct each stage: how much slower is DTF?
6. Compare DTF values to skewness values: are they related?

In other words, traditional DIF analysis under the DFIT approach fits focal and reference models and then can invoke IPR to determine the appropriate cut-offs for the DTF statistic to determine whether the NCDIF values of items should be examined to prune items in a stepwise procedure. This is computationally complex, expensive, and presents many opportunities for error. In practice, the first run of a model might inform whether more analysis is done, so the study may constrain itself to this first pass. By contract, examining the Mahalanobis distances of parameters of a single model for all examinees is procedurally straightforward and

computationally simple. Thus, the study proposes to compare DTF and the skewness of Mahalanobis distances of standard errors of parameter estimates. If they are found to be correlated, then the cut-off values established via the former statistic may have more attainable analogues in the latter. This could provide a faster way to invoke DFIT or, more generally, provide a deterministic method to suspect DIF that is more resilient and easier to compute.

Software

Uyar and Ozturk Gubes (2020) compared the accuracy of BILOG-MG, Mplus, and R's *ltm* for performing Maximum Likelihood Estimation of a 2PL under varying sample sizes and test lengths; they concluded that, despite observable differences, all three software options were viable choices. Given this outcome and its price, R was used for all aspects of the study: as free software, R would be most readily available to any evaluator. The ltm, DFIT, and slurmR packages facilitated the study.

The R workspace created by the simulation was large; specific variables were extracted to a smaller dataset during the supercomputer simulation process. The study intended to compare the DTF statistic to a new statistic based on implementation of the Mahalanobis distance on standard errors of parameter estimates. Variables extracted from models for analysis thus include DTF and the Mahalanobis-distance statistic. To analyze the theoretical relationship between these diagnostic statistics, item-level parameter estimates were extracted. For the practical aspects of the study, times to calculate these models and statistics were also extracted.

Data Cleaning

The original simulation produced 90,000 sets of trial data with three IRT models apiece from each of 625 repetitions through 120 permutations of four condition types, but these repetitions included small sample sizes with locally independent, uniform sampling in the

simulation of test responses. This resulted in the presence of generated test response data that would be perceived as unlikely and estimations that, consequentially, appear as outlying parameter estimates: some cleaning was warranted prior to analysis to focus analysis to trends within responses that would likely be retained for analysis in the field. A cleaned, simplified dataset was prepared for analysis based on identifying outliers via parameter estimates for the full model. A standardized statistic was calculated similar to root mean standard error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n_i} \frac{(\hat{x}_i - x_i)^2}{n_i}} \tag{31}$$

whose value is not unlike an averaged Euclidean distance of each model's item estimations from their true values. To account for multiple parameter traits and varying dimensionalities, this Euclidean distance was extended to incorporate both parameters' standardized differences as dimensions then comparing to a cut-off value of 4 in all dimensions for each trial's full model, t, with respect to its condition, c, roughly such that with each item, i:

$$Distance = \sqrt{\sum_{i=1}^{n_i} \left(\frac{\hat{b}_{t,i} - \bar{\hat{b}}_{t,i}}{\sigma_{\left(\hat{b}_{c,i} - b_{c,i}\right)}} \right)^2} + \sqrt{\sum_{i=1}^{n_i} \left(\frac{\hat{a}_{t,i} - \bar{\hat{a}}_{t,i}}{\sigma_{\left(\hat{a}_{c,i} - a_{c,i}\right)}} \right)^2} \tag{32}$$

In other words, since the measures of interest were instrument-wide, trials with models containing outliers were algorithmically identified as follows:

- Transform each trial's item parameter estimates to differences from true values
- Render differences into z-scores based on distribution across the condition
- Express estimates as Euclidean distances based on parameter type and DIF presence, resulting in 4 variables: difficulty/discrimination, DIF/no DIF
- Prepare cut-offs based on Euclidean distances equivalent to z=4 in all dimensions

For example, in trial 23 out of 90,000, the values for the process were:

Table 4

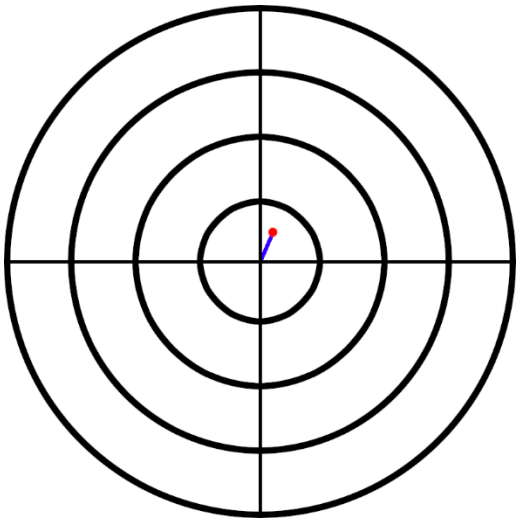
A Full Model's Standardized Parameter Estimates for Data Cleaning, Trial 23

Item	Estimate		True Value		Difference		Standardized	
	B	A	B	A	B	A	B	A
1	0.41	1.03	-0.07	0.49	0.48	0.54	0.18	0.45
2	0.54	2.77	0.21	0.92	0.33	1.85	0.20	2.65
3	0.66	2.00	0.54	1.26	0.12	0.74	0.21	1.68
4	0.13	0.81	-0.03	0.61	0.16	0.20	0.13	0.17
5	0.18	4.16	0.01	1.74	0.17	2.42	0.14	4.41
6	1.87	1.17	1.96	0.50	-0.09	0.67	0.39	0.63
7	-0.14	1.13	0.04	0.96	-0.18	0.17	0.10	0.58
8	0.29	1.31	-0.09	0.59	0.38	0.72	.016	0.80
9	-0.83	0.92	-1.16	0.82	0.33	0.10	-0.01	0.32
10	0.05	23.29	0.02	1.26	0.03	22.03	0.12	28.55
11	0.34	1.78	0.20	0.82	0.14	0.96	0.16	1.40
12	-0.53	0.85	-0.43	0.75	-0.10	0.10	0.04	0.23
13	-0.10	2.55	-0.06	1.49	-0.04	1.06	0.10	2.38
14	-0.35	3.97	-0.34	0.97	-0.01	3.00	0.06	4.16
15	0.36	1.45	0.05	1.49	0.31	-0.03	0.17	0.99

To proceed, first conceive of a dartboard for each item with rings representing a z-value of 1 for either parameter, resulting in something not unlike the following for Item 1:

Figure 8.

Depiction of Estimation Error and Euclidean Distance



The blue line represents the Euclidean distance between (0,0) and (0.18, 0.45) and is equal to 0.48. Now, extend this to a dartboard not for a single item, but a single model's parameters, creating four multidimensional dartboards for every DIF-free difficulty parameter, every DIF-free discrimination parameter, every DIF-containing difficulty parameter, and every DIF-containing discrimination parameter. Although it is quite beyond humans to conceive of high-dimensional dartboards, R can do it with the *dist* function:

```
dist(rbind(
  c(0.45, 2.65, 1.68, 0.17, 4.41,
    0.63, 0.58, 0.8, 0.32, 28.55,
    1.4, 0.23, 2.38, 4.16, 0.99),
  rep(0, times = 15)
))
```

The resulting value is 29.53, which can be compared to a cut-off value of 4 in all dimensions:

```
dist(rbind(
  rep(4, times = 15),
  rep(0, times = 15)
))
```

The resulting value is 15.42; trial 23's DIF-free discrimination parameter estimates were thus among those found to be so extreme as to be problematic. A researcher looking at a discrimination estimate of 28.55 for Item 10 would likely come to the same conclusion and omit it from subsequent analysis. With this deterministic algorithm defined, the cut-offs can be evaluated for each trial's full model by the computer. The cut-off process rejected approximately 7.5% of all simulation trials from the analysis, spread among study conditions as follows:

Table 5

Trials Omitted from Analysis based on Outlying Estimates (6,752 rejections total)

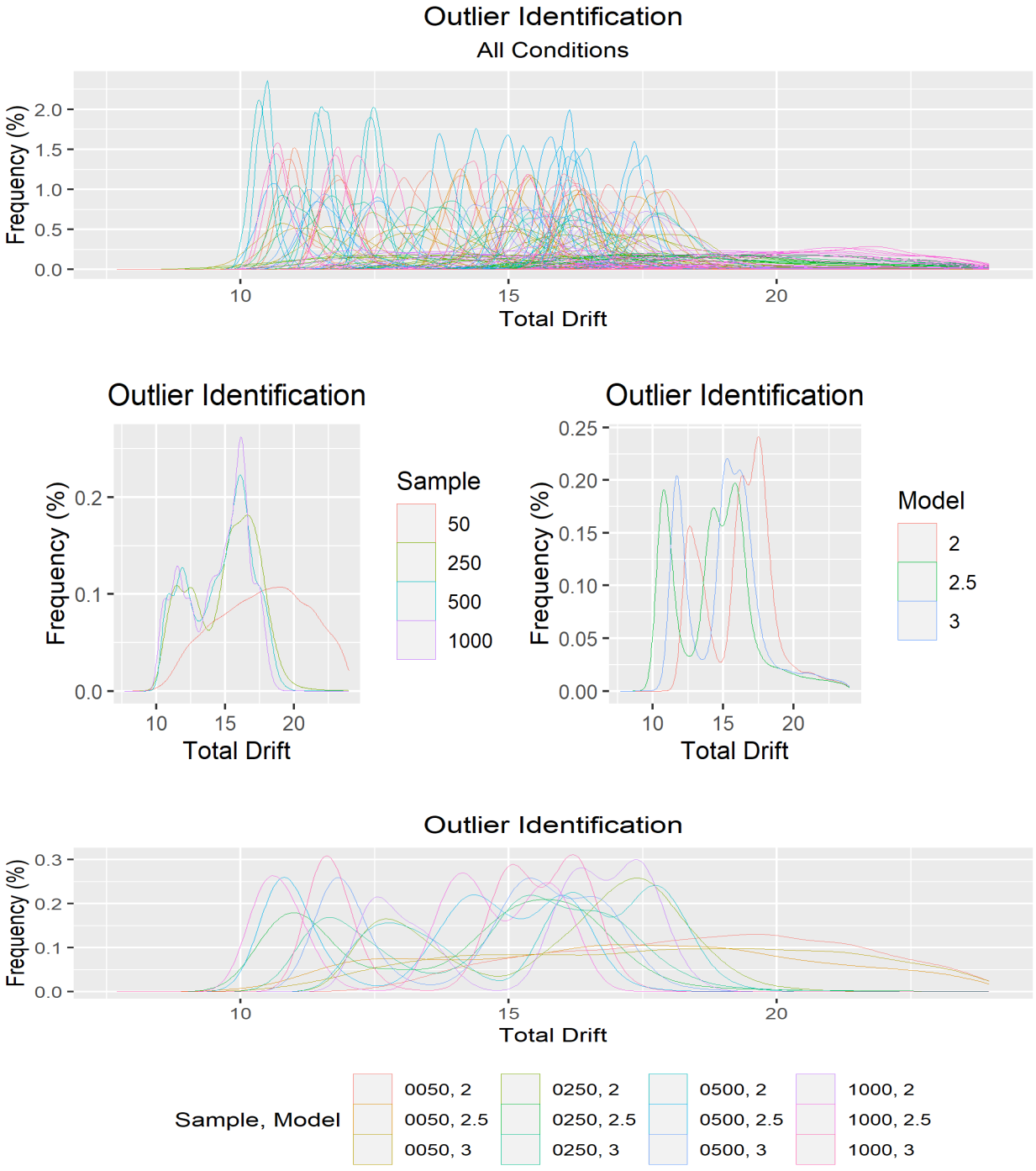
Condition	Omissions
Test Length	
15	3,184
30	3,568
Sample Size	
50	6,705
250	44
500	3
1000	0
DIF Presence	
(None)	3,136
20% @ -0.5	1,030
20% @ -1	1,072
33% @ -0.5	1,148
33% @ -1	1,366
Generating Model	
2PL	3,080
2PL+C	1,680
3PL	1,992

Outlier Distribution

The small sample size condition clearly bears responsibility for most of the outliers, but a disproportionate amount of 2PL models were also observed to present problems. Figure 9 depicts the situation visually with density plots of the total drift—the sum of all four unidimensional Euclidean distances of parameter estimation differences of all items described above—under varying conditions. Thus, not only does the N=50 condition fail to converge more than any other, but the models that do manage to converge are full of more bad estimates than any other model!

Figure 9

Outlier Identification under Various Condition Groupings



4 RESULTS

The simulation study produced outcome variables in three categories: the parameter estimates for models, the DIF diagnostics including both DTF and the Mahalanobis distance-based statistic (MD), and the time cost of models and measures. The simulation was conducted on a supercomputer allocation of 48 threads for a total of 18 days, 12 hours of CPU time with a maximum usage of 194 gigabytes of memory, but some models did not converge.

Non-Convergence Errors

Using the tryCatch() R method, the simulation was able to gracefully recover from errors and record where they occurred throughout the study. This allowed for recording frequencies:

Table 6.

Convergence and Evaluation Error Frequencies, by condition (out of 90,000 trials)

Condition	Models			Statistics[1]	
	Full	Reference	Focal	DTF	MD
Test Length					
15	15	34	1,567	7,254	23
30	4	21	5,079	5,711	23
Sample Size					
50	19	55	6,635	11,363	46
250	0	0	11	1,362	0
500	0	0	0	212	0
1000	0	0	0	28	0
DIF Presence					
(None)	14	15	1,229	6,470	34
20% @ -0.5	1	7	650	3,156	16
20% @ -1	1	19	1,436	3,128	11
33% @ -0.5	2	3	1,208	3,424	14
33% @ -1	1	11	2,123	3,483	11
Generating Model					
2PL	16	26	1,933	3,058	28
2PL+C	3	15	1,986	5,743	5
3PL	0	14	2,727	4,161	13

[1] 83,296 trials had all models fit and were suited for statistics

As the anecdote goes, “the absence of data is data,” and observing which conditions presented problems with model convergence and statistic evaluation empirically confirmed that small sample sizes make IRT difficult: the majority of all problems occurred in the smallest sample size condition. Within convergent trials, DTF calculation failed over 400% more frequently than the Mahalanobis-based statistic.

In the study’s fitted IRT models, the focal group model for the small sample size condition is tremendously small, N=12, and below acceptable levels for IRT analysis; estimating parameters for this group with the additional challenge of constraints from the reference group presented a scenario in which failure was expected, but its amount was unknown. This small-sized condition was chosen based on real-world school sizes to empirically reveal not *if* the models would break—such an outcome is to be expected—but how often and how badly those failures occurred. The convergence failures are high, but they represent the least problematic error type; the error screen an everyday researcher would be presented with is difficult to misinterpret as a successful analysis. Within successful convergence, however, DTF statistics failed to calculate almost twice as often in this condition.

An examination of randomly-selected error conditions revealed the cause: often, in these poorly-estimated models, the discrimination parameter, *a*, was erroneously placed at a negative level. This causes DTF measurement to fail using the *DFIT* package due to mathematical constraints not unlike those present in Raju’s area method when pseudo-random (*c*) parameters are unequal; the resulting curves do not conform to the calculus which estimates their areas.

Overall Findings

Simulation findings, including parameter estimates and DIF statistics, were examined by condition, with the overall outcomes within each condition level. The correlation between the

traditional DIF measure of DTF and the new, proposed DIF measure based on the Mahalanobis distances of parameter estimation standard errors, was calculated within each condition permutation. This allowed for the influence of each combination of variables to be undistorted by amalgamation within that group. Additionally, the number of degrees of freedom available for that correlation in turn allows for comparison of the availability of DTF within each scenario. In several scenarios, mostly situated within the N=50 conditions, the DTF statistic could not be calculated often enough to attempt a correlation. For the DIF presence condition, composed of a combination of DIF magnitude and DIF percent, some permutations of the simulation combined to create a fully factorial design, but the original data generation included unequal groups as a result. For these cases, a random subsample of 625 trials was selected from available data to provide comparable sample sizes for analysis.

The outcomes are summarized in Table 7 by condition level and include parameter estimation information in the form of RMSE, the DIF measurement statistics, and the time it took for each condition to provide these various outcome variables. Table 8 goes into more detail per condition and includes the correlations between DIF measurement statistics. In all these tables, generally speaking, values closer to 0 are better except in cases where DIF should be measured and for correlation statistics. Correlation is observed varying through permutations of simulations conditions, suggesting that the relationship between the DIF measurement statistics may vary depending on those conditions. Further analysis were conducted to explore these potential relationships. Additionally, general trends in RMSE clearly demonstrate the benefits of additional data in the form of sample size and test length.

Table 7.

Overall Simulation Outcomes by Condition

Condition	Parameter RMSE M(SD)		DTF		MD		Seconds to Compute M(SD)		
	b	*a*	M	(SD)	M	(SD)	Full Model	DTF	MD[1]
Test Length									
15	0.70 (0.625)	1.00 (0.791)	0.68	(1.026)	2.19	(0.433)	0.82 (0.400)	2.25 (0.331)	0.54 (2.048)
30	0.68 (0.619)	0.94 (0.759)	1.93	(2.703)	3.64	(0.787)	2.66 (1.445)	8.42 (1.076)	0.55 (0.252)
Sample Size									
50	0.95 (0.924)	1.34 (1.066)	3.31	(4.518)	2.67	(0.863)	0.59 (0.282)	4.33 (3.287)	0.54 (0.232)
250	0.65 (0.537)	0.91 (0.674)	1.55	(2.177)	2.81	(0.967)	1.13 (0.565)	4.66 (2.752)	0.54 (0.251)
500	0.63 (0.506)	0.87 (0.658)	1.11	(1.832)	2.92	(0.965)	1.79 (0.943)	5.10 (2.965)	0.54 (0.223)
1000	0.61 (0.482)	0.86 (0.651)	0.96	(1.585)	3.16	(0.965)	3.09 (1.697)	6.15 (3.535)	0.56 (2.786)
DIF Presence									
(None)	0.97 (0.810)	0.68 (0.611)	0.52	(1.404)	2.91	(0.966)	1.74 (1.400)	5.19 (3.120)	0.54 (0.213)
20% @ -0.5	0.97 (0.780)	0.68 (0.619)	0.57	(1.154)	2.92	(0.969)	1.72 (1.395)	5.27 (3.155)	0.53 (0.148)
20% @ -1	0.98 (0.779)	0.68 (0.615)	1.02	(1.596)	2.94	(0.972)	1.74 (1.394)	5.27 (3.192)	0.54 (0.146)
33% @ -0.5	0.96 (0.738)	0.71 (0.641)	1.34	(1.975)	2.85	(0.944)	1.74 (1.420)	5.27 (3.225)	0.57 (3.538)
33% @ -1	0.95 (0.734)	0.71 (0.634)	3.83	(2.694)	2.90	(0.959)	1.75 (1.403)	5.44 (3.280)	0.54 (0.455)
Generating Model									
2PL	0.58 (0.472)	1.23 (0.884)	1.68	(2.640)	3.04	(0.926)	1.76 (1.416)	5.04 (3.089)	0.55 (2.544)
2PL+C	0.76 (0.702)	0.76 (0.623)	1.10	(1.885)	2.80	(0.984)	1.71 (1.387)	5.53 (3.282)	0.54 (0.265)
3PL	0.72 (0.648)	0.93 (0.731)	1.07	(1.619)	2.89	(0.960)	1.74 (1.403)	5.27 (3.165)	0.54 (0.264)

Note. RMSE = Root Mean Square Error; DTF = Differential Test Functioning Statistic; MD = Mahalanobis Distance-based Statistic

[1] Time for Mahalanobis Distance-based Statistics given in milliseconds; 0.54 milliseconds = 54/1000 of a second

Table 8.

Simulation Findings and Correlation Between DTF and MD by Permutation

N_i	Model	DIF	N	RMSE		DTF		MD		Correlation		
				b	*a*	M	(SD)	M	(SD)	*df*	*r*	*p(.05)*
15	2PL	(None)	50	0.71 (0.643)	1.87 (1.360)	1.96	(1.725)	2.04	(0.383)	391	.01	.828
			250	0.58 (0.428)	1.24 (0.823)	0.30	(0.305)	2.24	(0.348)	619	-.07	.093
			500	0.54 (0.400)	1.29 (0.872)	0.14	(0.127)	2.33	(0.269)	623	-.07	.084
			1,000	0.53 (0.385)	1.23 (0.834)	0.06	(0.060)	2.41	(0.211)	623	-.06	.152
		20% @ -0.5	50	0.78 (0.780)	1.47 (1.073)	1.84	(1.656)	2.01	(0.378)	268	-.03	.631
			250	0.56 (0.419)	1.26 (0.849)	0.35	(0.284)	2.2	(0.364)	608	0	.950
			500	0.56 (0.407)	1.22 (0.823)	0.11	(0.106)	2.31	(0.276)	623	0	.998
			1,000	0.55 (0.398)	1.19 (0.808)	0.11	(0.089)	2.4	(0.201)	623	.03	.524
		20% @ -1	50	0.78 (0.724)	1.55 (1.078)	3.99	(3.303)	2.05	(0.366)	112	-.03	.755
			250	0.54 (0.406)	1.38 (0.931)	0.47	(0.417)	2.23	(0.285)	621	-.02	.689
			500	0.55 (0.409)	1.21 (0.818)	0.28	(0.225)	2.26	(0.29)	623	-.05	.177
			1,000	0.54 (0.391)	1.24 (0.841)	0.18	(0.133)	2.36	(0.194)	623	.01	.879
		33% @ -0.5	50	0.86 (0.772)	1.40 (0.985)	2.40	(1.912)	1.98	(0.379)	107	-.28	.004
			250	0.57 (0.434)	1.24 (0.837)	1.04	(0.682)	2.25	(0.375)	618	0	.953
			500	0.55 (0.406)	1.24 (0.834)	0.49	(0.303)	2.38	(0.247)	622	.06	.158
			1,000	0.56 (0.410)	1.17 (0.796)	0.77	(0.286)	2.41	(0.204)	623	.07	.075
		33% @ -1	50	0.77 (0.793)	1.52 (1.129)	4.28	(3.232)	2.06	(0.385)	246	.01	.891
			250	0.63 (0.481)	1.15 (0.794)	2.71	(1.025)	2.26	(0.364)	605	0	.995
			500	0.59 (0.436)	1.17 (0.797)	2.79	(0.754)	2.39	(0.291)	621	.02	.541
			1,000	0.55 (0.408)	1.19 (0.811)	2.20	(0.451)	2.47	(0.215)	623	.06	.105
	2PL+C	(None)	50	1.06 (0.980)	1.07 (0.858)	1.57	(1.490)	1.98	(0.354)	43	.37	.012
			250	0.76 (0.641)	0.63 (0.423)	0.22	(0.206)	2.03	(0.492)	564	.05	.213
			500	0.68 (0.567)	0.67 (0.454)	0.12	(0.116)	2.07	(0.515)	601	.02	.594
			1,000	0.66 (0.527)	0.64 (0.459)	0.09	(0.077)	2.22	(0.427)	621	.03	.465
		20% @ -0.5	50	1.18 (1.070)	1.12 (0.855)	1.46	(1.627)	2.00	(0.380)	32	.25	.149
			250	0.74 (0.671)	0.63 (0.430)	0.19	(0.181)	2.13	(0.463)	524	-.01	.754

Table 8.

(Continued)

N_i	Model	DIF	N	RMSE		DTF		MD		Correlation		
				b	*a*	M	(SD)	M	(SD)	*df*	*r*	*p(.05)*
15	2PL+C	20% @ -0.5	500	0.67 (0.542)	0.65 (0.455)	0.54	(0.326)	2.09	(0.505)	605	-.02	.644
			1,000	0.68 (0.551)	0.65 (0.453)	0.08	(0.073)	2.13	(0.473)	620	.05	.205
		20% @ -1	50	1.07 (1.006)	1.27 (1.036)	1.76	(1.585)	1.95	(0.407)	121	.14	.110
			250	0.70 (0.588)	0.80 (0.559)	0.55	(0.387)	2.08	(0.492)	568	-.06	.129
			500	0.70 (0.566)	0.62 (0.430)	0.22	(0.161)	2.12	(0.495)	610	0	.932
			1,000	0.68 (0.555)	0.63 (0.438)	0.20	(0.127)	2.18	(0.455)	620	-.03	.445
		33% @ -0.5	50	0.92 (0.869)	1.13 (0.854)	1.43	(1.357)	2.02	(0.370)	146	-.01	.933
			250	0.73 (0.617)	0.74 (0.516)	0.45	(0.336)	1.99	(0.530)	566	-.03	.427
			500	0.72 (0.585)	0.66 (0.456)	0.64	(0.298)	2.06	(0.533)	613	-.01	.773
			1,000	0.67 (0.543)	0.67 (0.467)	0.47	(0.189)	2.18	(0.457)	621	.06	.111
		33% @ -1	50	1.12 (0.949)	1.11 (0.884)	1.31	(1.087)	1.97	(0.39)	53	.01	.957
			250	0.80 (0.684)	0.73 (0.501)	2.29	(0.863)	1.97	(0.545)	520	-.01	.846
			500	0.70 (0.571)	0.68 (0.477)	1.78	(0.549)	2.14	(0.482)	614	.02	.677
			1,000	0.68 (0.549)	0.65 (0.460)	1.25	(0.329)	2.22	(0.456)	617	.03	.505
	3PL	(None)	50	0.94 (0.964)	1.34 (1.131)	1.48	(1.492)	1.99	(0.365)	79	-.09	.401
			250	0.66 (0.541)	0.92 (0.650)	0.31	(0.335)	2.11	(0.489)	575	.05	.270
			500	0.65 (0.514)	0.89 (0.609)	0.10	(0.090)	2.21	(0.482)	612	0	.984
			1,000	0.65 (0.494)	0.83 (0.584)	0.07	(0.065)	2.40	(0.366)	623	-.03	.520
		20% @ -0.5	50	1.11 (1.071)	1.17 (1.020)	1.27	(1.179)	2.02	(0.343)	54	.09	.523
			250	0.64 (0.514)	0.92 (0.630)	0.18	(0.182)	2.10	(0.489)	596	.03	.450
			500	0.70 (0.553)	0.80 (0.554)	0.08	(0.072)	2.17	(0.515)	610	.05	.238
			1,000	0.64 (0.490)	0.84 (0.588)	0.07	(0.069)	2.51	(0.294)	622	.06	.141
		20% @ -1	50	0.75 (0.818)	1.52 (1.120)	2.42	(1.922)	2.03	(0.364)	266	.06	.355
			250	0.65 (0.516)	0.90 (0.627)	0.34	(0.253)	2.10	(0.496)	579	-.01	.796
			500	0.63 (0.491)	0.88 (0.614)	0.23	(0.178)	2.36	(0.388)	613	0	.974
			1,000	0.66 (0.515)	0.84 (0.588)	0.45	(0.186)	2.39	(0.370)	623	.08	.057
		33% @ -0.5	50	0.89 (0.842)	1.21 (0.879)	1.28	(1.196)	1.98	(0.370)	139	-.07	.432

Table 8.

(Continued)

N_i	Model	DIF	N	RMSE		DTF		MD		Correlation		
				b	*a*	M	(SD)	M	(SD)	*df*	*r*	*p(.05)*
15	3PL	33% @ -0.5	250	0.75 (0.613)	0.88 (0.611)	0.81	(0.526)	2.04	(0.497)	596	-.02	.644
			500	0.64 (0.502)	0.89 (0.620)	0.65	(0.324)	2.36	(0.352)	619	-.05	.175
			1,000	0.68 (0.527)	0.81 (0.561)	0.44	(0.194)	2.43	(0.333)	621	.03	.420
15	3PL	33% @ -1	50	0.90 (0.912)	1.39 (0.943)	3.23	(2.701)	1.99	(0.400)	62	-.05	.678
			250	0.74 (0.610)	0.77 (0.535)	1.61	(0.725)	2.09	(0.500)	598	.06	.149
			500	0.74 (0.578)	0.75 (0.518)	1.91	(0.571)	2.19	(0.469)	620	.02	.671
			1,000	0.67 (0.520)	0.81 (0.570)	1.24	(0.300)	2.42	(0.331)	621	-.03	.427
30	2PL	(None)	50	0.67 (0.571)	1.72 (1.282)	7.98	(8.337)	3.37	(0.692)	153	-.04	.658
			250	0.57 (0.411)	1.08 (0.746)	1.77	(1.939)	3.68	(0.693)	597	.01	.836
			500	0.53 (0.396)	1.10 (0.761)	0.64	(0.814)	3.91	(0.567)	623	.15	0
			1,000	0.51 (0.377)	1.06 (0.745)	0.37	(0.494)	4.15	(0.354)	623	-.03	.477
		20% @ -0.5	50	0.74 (0.757)	1.67 (1.185)	4.59	(4.263)	3.36	(0.695)	129	-.11	.223
			250	0.57 (0.424)	1.01 (0.696)	1.01	(1.136)	3.75	(0.640)	612	-.02	.556
			500	0.53 (0.396)	1.06 (0.733)	0.42	(0.506)	3.91	(0.541)	622	-.02	.655
			1,000	0.50 (0.367)	1.12 (0.776)	0.87	(0.654)	4.21	(0.303)	623	0	.918
		20% @ -1	50	0.74 (0.683)	1.49 (1.077)	6.06	(6.590)	3.37	(0.695)	95	.03	.736
			250	0.53 (0.403)	1.17 (0.800)	3.25	(2.814)	3.69	(0.669)	613	.01	.756
			500	0.55 (0.406)	1.00 (0.700)	1.69	(1.007)	3.88	(0.546)	621	-.06	.115
			1,000	0.53 (0.386)	1.03 (0.722)	1.49	(0.822)	4.15	(0.305)	623	-.04	.382
		33% @ -0.5	50	0.68 (0.580)	1.61 (1.112)	14.83	(15.874)	3.34	(0.702)	41	.19	.220
			250	0.55 (0.417)	1.14 (0.779)	4.28	(3.106)	3.58	(0.734)	604	.04	.317
			500	0.53 (0.396)	1.10 (0.764)	2.44	(1.770)	3.91	(0.589)	623	.06	.147
			1,000	0.52 (0.391)	1.09 (0.756)	1.87	(1.040)	4.11	(0.365)	623	-.04	.295
		33% @ -1	50	0.69 (0.617)	1.62 (1.067)	—	—	3.28	(0.761)	—	—	—
			250	0.58 (0.438)	1.04 (0.718)	5.51	(2.966)	3.63	(0.707)	593	-.06	.166

Table 8.

(Continued)

N_i	Model	DIF	N	RMSE		DTF		MD		Correlation		
				b	*a*	M	(SD)	M	(SD)	*df*	*r*	*p(.05)*
30	2PL	33% @ -1	500	0.53 (0.398)	1.14 (0.783)	7.68	(3.082)	3.87	(0.588)	623	0	.950
			1,000	0.56 (0.423)	1.00 (0.701)	6.63	(1.666)	4.16	(0.406)	623	-.02	.567
30	2PL+C	(None)	50	0.95 (0.901)	1.21 (1.055)	5.44	(4.179)	3.34	(0.642)	37	-.11	.493
			250	0.67 (0.552)	0.67 (0.460)	0.75	(0.759)	3.29	(0.904)	574	-.02	.646
			500	0.64 (0.541)	0.63 (0.438)	0.33	(0.329)	3.62	(0.827)	611	-.09	.029
			1,000	0.59 (0.498)	0.60 (0.423)	0.12	(0.128)	3.91	(0.709)	623	.05	.191
		20% @ -0.5	50	0.78 (0.723)	1.23 (1.064)	4.67	(4.348)	3.36	(0.717)	107	.05	.602
			250	0.63 (0.525)	0.72 (0.503)	0.77	(0.803)	3.54	(0.859)	569	0	.949
			500	0.68 (0.558)	0.65 (0.460)	0.3	(0.338)	3.40	(0.926)	615	-.04	.358
			1,000	0.64 (0.531)	0.64 (0.455)	0.36	(0.301)	3.80	(0.752)	623	-.04	.379
		20% @ -1	50	1.08 (1.012)	1.02 (0.874)	—	—	3.38	(0.601)	—	—	—
			250	0.69 (0.579)	0.72 (0.516)	1.73	(1.446)	3.27	(0.926)	568	.02	.569
			500	0.63 (0.520)	0.64 (0.462)	1.02	(0.743)	3.49	(0.862)	619	-.03	.461
			1,000	0.63 (0.513)	0.57 (0.405)	0.87	(0.466)	4.04	(0.624)	621	-.04	.336
		33% @ -0.5	50	1.39 (1.256)	1.05 (0.845)	—	—	3.40	(0.592)	—	—	—
			250	0.78 (0.663)	0.69 (0.471)	1.57	(1.404)	3.16	(0.926)	547	-.04	.330
			500	0.72 (0.599)	0.65 (0.460)	0.77	(0.637)	3.27	(0.929)	609	-.03	.500
			1,000	0.66 (0.555)	0.69 (0.498)	1.09	(0.632)	3.40	(0.842)	621	-.07	.080
		33% @ -1	50	1.42 (1.206)	0.94 (0.720)	—	—	3.30	(0.630)	—	—	—
			250	0.65 (0.536)	0.72 (0.504)	7.51	(3.320)	3.54	(0.854)	569	-.03	.499
			500	0.70 (0.584)	0.68 (0.482)	5.39	(1.870)	3.45	(0.903)	613	-.08	.037
			1,000	0.65 (0.539)	0.67 (0.477)	5.11	(1.256)	3.85	(0.753)	623	-.02	.568
	3PL	(None)	50	1.03 (1.038)	1.42 (1.134)	5.42	(4.854)	3.40	(0.626)	74	-.01	.900
			250	0.66 (0.522)	0.81 (0.561)	1.06	(1.105)	3.60	(0.839)	567	.04	.309
			500	0.63 (0.518)	0.81 (0.555)	0.39	(0.486)	3.38	(0.949)	621	0	.984

Table 8.

(Continued)

N_i	Model	DIF	N	RMSE		DTF		MD		Correlation		
				b	*a*	M	(SD)	M	(SD)	*df*	*r*	*p(.05)*
30	3PL	(None)	1,000	0.61 (0.497)	0.80 (0.569)	0.15	(0.152)	3.83	(0.734)	623	.06	.110
		20% @ -0.5	50	0.95 (0.914)	1.59 (1.177)	3.35	(3.203)	3.31	(0.662)	43	-.08	.623
			250	0.65 (0.518)	0.83 (0.573)	1.37	(1.317)	3.57	(0.824)	606	.05	.217
			500	0.66 (0.526)	0.79 (0.551)	0.79	(0.667)	3.55	(0.864)	617	-.07	.074
			1,000	0.62 (0.498)	0.80 (0.565)	0.21	(0.209)	3.96	(0.658)	623	-.04	.372
		20% @ -1	50	1.17 (1.076)	1.23 (0.944)	6.86	(5.026)	3.39	(0.641)	16	0.1	.693
			250	0.66 (0.529)	0.82 (0.579)	1.49	(1.339)	3.35	(0.879)	594	-.03	.461
			500	0.63 (0.499)	0.81 (0.566)	0.74	(0.652)	3.76	(0.839)	619	0	.923
			1,000	0.60 (0.473)	0.78 (0.557)	0.78	(0.464)	4.2	(0.54)	623	.07	.067
		33% @ -0.5	50	1.18 (1.029)	1.23 (0.892)	—	—	3.28	(0.72)	—	—	—
			250	0.67 (0.542)	0.88 (0.607)	2.84	(2.176)	3.32	(0.912)	604	-.04	.348
			500	0.69 (0.555)	0.77 (0.531)	1.26	(0.939)	3.62	(0.876)	616	-.03	.411
			1,000	0.62 (0.492)	0.79 (0.554)	1.14	(0.657)	3.96	(0.628)	623	.06	.142
		33% @ -1	50	0.90 (0.833)	1.45 (0.982)	—	—	3.32	(0.663)	—	—	—
			250	0.65 (0.522)	0.86 (0.601)	4.4	(2.507)	3.48	(0.874)	603	.05	.181
			500	0.67 (0.541)	0.82 (0.576)	4.27	(1.78)	3.31	(0.929)	622	.02	.659
			1,000	0.65 (0.518)	0.75 (0.529)	4.75	(1.175)	3.82	(0.725)	622	.03	.511

Note. Ni = number of items, Model = generating model, DIF = DIF presence, N = sample size, b = difficulty, a = discrimination, DTF = differential test functioning statistic, MD = Mahalanobis distance-based statistic, Correlation = Pearson correlation between DTF and MD.

Parameter Estimation

Within the full model, parameter estimates exhibited varying accuracy both within and between conditions, perhaps because of both the influence of random number generation and conditions. The smallest conditions (N=50, N_i=15) had only 2.5% of the data that the largest conditions (N=1000, N_i=30) did, allowing for more influence from outliers arising from random number generation. Part of the simulation data generation uses uniform sampling distributions, potentially intensifying this effect. Additionally, one of the conditions with the most DIF but the least amount of data (N=50, N_i=30, 2PL, 33% DIF @ -1) demonstrated an intense difficulty in calculating; the entire condition struggled to produce DTF statistics to measure correlation. Appendix A presents detailed charts of the mean differences between estimated and actual parameters, and their standard deviations, across simulation conditions for all items.

To efficiently analyze parameter accuracy between conditions of varying test length, and to facilitate analyses of variance, a dimensionality reduction procedure using an unstandardized version of the adapted RMSE from the data cleaning procedures were used. For example, in trial 75,626, the parameter estimations were first rescaled as errors, or differences from the true values, then were first transformed into Z-scores based on findings within the condition (in this case, N_i = 30, N = 1,000, DIF = 0%, Magnitude = -0.5, Model = 2PLC) to create standardized measures of mean differences for data cleaning:

Table 9.

Example Estimator Simplification for Trial 75,626

Item	Trial Estimate		True Value		Difference		Standardized	
	B	A	B	A	B	A	B	A
1	-0.75	0.64	-0.07	0.49	-0.68	0.15	-0.25	-0.13
2	-0.32	1.06	0.21	0.92	-0.53	0.14	-1.63	0.13
3	0.29	1.09	0.54	1.26	-0.25	-0.17	1.63	-0.44
4	-0.71	0.67	-0.03	0.61	-0.68	0.06	-1.39	-1.40
5	-0.26	1.90	0.01	1.74	-0.27	0.16	1.11	0.93
6	1.47	0.52	1.96	0.50	-0.49	0.02	-0.64	1.54
7	-0.39	1.09	0.04	0.96	-0.43	0.13	-0.30	-0.65
8	-0.41	0.80	-0.09	0.59	-0.32	0.21	2.23	0.37
9	-1.55	1.15	-1.16	0.82	-0.39	0.33	-0.47	-0.77
10	-0.41	1.37	0.02	1.26	-0.43	0.11	-1.20	-0.45
11	-0.3	0.96	0.20	0.82	-0.50	0.14	-0.63	-0.05
12	-0.86	1.06	-0.43	0.75	-0.43	0.31	0.14	0.24
13	-0.43	1.89	-0.06	1.49	-0.37	0.40	-0.67	1.83
14	-0.85	1.12	-0.34	0.97	-0.51	0.15	-1.78	-1.45
15	-0.29	1.74	0.05	1.49	-0.34	0.25	-0.12	1.17
16	-0.72	1.23	-0.25	0.89	-0.47	0.34	-0.85	0.61
17	-0.28	1.66	0.06	1.45	-0.34	0.21	0.07	1.04
18	-0.11	0.83	0.31	0.75	-0.42	0.08	0.73	-0.47
19	-0.27	1.60	0.04	1.43	-0.31	0.17	0.64	0.42
20	-0.33	0.83	0.13	0.60	-0.46	0.23	1.02	0.91
21	0.09	0.88	0.52	0.83	-0.43	0.05	0.08	-0.08
22	-1.47	0.81	-0.96	0.56	-0.51	0.25	0.09	-0.02
23	-1.16	1.03	-0.79	0.67	-0.37	0.36	0.67	0.61
24	-0.17	0.68	0.37	0.7	-0.54	-0.02	-0.42	-1.62
25	-0.99	1.53	-0.71	1.03	-0.28	0.50	0.73	0.63
26	-0.67	1.13	-0.19	0.89	-0.48	0.24	-1.02	-0.15
27	0.41	1.09	0.74	1.23	-0.33	-0.14	-0.19	0.92
28	-0.81	1.27	-0.44	0.90	-0.37	0.37	0.15	0.55
29	-0.54	1.45	-0.17	1.23	-0.37	0.22	-0.30	-0.34
30	-0.04	0.80	0.53	0.69	-0.57	0.11	-0.80	0.20

Thus, each standardized parameter (in this case, DIF-free difficulty and DIF-free estimation) was expressed as a Euclidean distance of 30-dimensional vectors and summed, giving $Distance_{(Standardized)} = 9.64$, within range for retention in the dataset. Next, for analysis, those raw difference, or estimation errors, are instead amalgamated into Euclidean distances for each parameter for all items, *i*, in trial *t*:

$$Distance\,(B+A) = \sqrt{\sum_{i=1}^{n_i}\left(\hat{b}_{t,i} - b_{t,i}\right)^2} + \sqrt{\sum_{i=1}^{n_i}\left(\hat{a}_{t,i} - a_{t,i}\right)^2} \quad (33)$$

This allows for the estimation of the entire instrument to be expressed as a single number. Thus, the 30 DIF-free difficulty parameters and 30 DIF-free discrimination parameters for trial *75,626* are simplified to a sum of two Euclidean distances each measured in 30 orthogonal dimensions, or $Distance_{(Difference)} = 3.69$. This allowed for comparing conditions with a single variable:

Table 10

Parameter Estimation Accuracy by Condition, M(SD)

Condition	Parameter RMSE M(SD)		Simplified Distance	
	b	*a*	M	(SD)
Test Length				
15	0.70 (0.625)	1.00 (0.791)	4.13	(2.191)
30	0.68 (0.619)	0.94 (0.759)	5.81	(3.226)
Sample Size				
50	0.95 (0.924)	1.34 (1.066)	9.05	(4.458)
250	0.65 (0.537)	0.91 (0.674)	4.33	(1.016)
500	0.63 (0.506)	0.87 (0.658)	3.98	(0.788)
1000	0.61 (0.482)	0.86 (0.651)	3.71	(0.664)
% DIF Items				
0	0.68 (0.611)	0.97 (0.810)	4.89	(3.046)
20%	0.68 (0.617)	0.97 (0.780)	4.95	(2.873)
33%	0.71 (0.638)	0.95 (0.736)	5.06	(2.708)
Magnitude				
-0.5	0.70 (0.630)	0.97 (0.778)	4.96	(2.950)
-1	0.68 (0.614)	0.96 (0.773)	4.97	(2.809)
Generating Model				
2PL	0.58 (0.472)	1.23 (0.884)	4.78	(2.229)
2PL+C	0.76 (0.702)	0.76 (0.623)	5.13	(3.146)
3PL	0.72 (0.648)	0.93 (0.731)	4.97	(3.133)

Another, more ordinal way of examining these accuracies was to sort by these simplified distances and examine which conditions are associated with the lowest and highest values. As expected from the above averages, the largest sample size condition is one of the most consistently accurate:

Table 11

Best and worst estimated full models by average parameter fit

N_i	N	DIF	Magnitude	Model
30	1,000	0	-0.5	2PLC
30	1,000	0	-1	2PLC
30	500	0	-0.5	2PLC
30	500	0	-1	2PLC
30	250	0	-0.5	2PLC
15	1,000	0	-0.5	2PLC
15	1,000	0	-1	2PLC
15	500	0	-0.5	2PLC
15	500	0	-1	2PLC
30	1,000	0	-1	3PL

N_i	N	DIF	Magnitude	Model
15	50	33%	-1	3PL
15	50	33%	-0.5	3PL
30	50	33%	-1	2PLC
30	50	33%	-0.5	3PL
30	50	20%	-1	3PL
30	50	33%	-0.5	2PLC
15	50	20%	-1	2PL
15	50	33%	-1	3PL
15	50	33%	-0.5	3PL
30	50	33%	-1	2PLC

A more traditional analysis of variance between condition groups was used to identify which conditions were associated with significant differences in parameter estimates. A three-way analysis of variance demonstrated the problem of large sample sizes: all factors and interactions demonstrated statistical significance. Practical significance was examined using η^2 and showed test length, sample size, and the interaction between the two as practically significant. Using traditional effect size cut-off values of .01, .06, and .14, the interaction effect displayed small practical significance, test length a medium, and sample size a large impact:

Table 12

Analysis of Variance for Full Model's Simplified Parameter Estimate

Condition	Sum of Squares	*df*	Mean Square	*F*	*p(F)*	η^2
Test Length	58,713	1	58,713	18,224.31	< .01	.09
Sample Size	332,758	3	110,919	34,428.73	< .01	.48
DIF Presence	901	4	225	69.90	< .01	< .01
Generating Model	278	2	139	43.17	< .01	< .01
Length x Size	17,524	3	5,841	1,813.09	< .01	.03
Length x DIF	1,019	4	255	78.08	< .01	< .01
Length x Model	1,226	2	613	190.23	< .01	< .01
Size x DIF	959	12	80	24.80	< .01	< .01
Size x Model	3,203	6	534	165.72	< .01	< .01
DIF x Model	1,499	8	187	58.17	< .01	< .01
Length x Size x DIF	1,002	12	83	25.91	< .01	< .01
Length x Size x Model	1,343	6	224	69.47	< .01	< .01
Length x DIF x Model	829	8	104	32.15	< .01	< .01
Size x DIF x Model	1,446	24	60	18.70	< .01	< .01

The very high sample sizes presented by this simulation undermine some of the interpretability of an analysis of variance, because statistical significant is more assured from sample size alone. Analyzing practical significance allowed for more focus. Based on this analysis, an examination of permutations between test length and sample size conditions was conducted:

Table 13.

Simulation Outcomes for Test Length (N_i) and Sample Size (N) Conditions

Condition	Simplified Distance		DTF		MD	
	M	(SD)	M	(SD)	M	(SD)
$N_i = 15$						
$N = 50$	7.30	(3.302)	2.25	(2.229)	2.01	(0.376)
$N = 250$	3.68	(0.819)	0.70	(0.869)	2.12	(0.464)
$N = 500$	3.30	(0.393)	0.58	(0.817)	2.22	(0.438)
$N = 1{,}000$	3.12	(0.258)	0.44	(0.601)	2.34	(0.366)
$N_i = 30$						
$N = 50$	10.91	(4.759)	6.28	(7.202)	3.36	(0.665)
$N = 250$	4.98	(0.740)	2.40	(2.696)	3.50	(0.837)
$N = 500$	4.67	(0.404)	1.64	(2.340)	3.62	(0.829)
$N = 1{,}000$	4.30	(0.353)	1.47	(2.033)	3.97	(0.637)

Some of the impact of test length on the unstandardized, unidimensional parameter estimation distance used for analysis is natural: more sampling creates more opportunities for distance, not unlike how the mean of a χ^2 distribution behaves as its *k* (degrees of freedom) value increases. By examining both test length and sample size conditions at the same time, this natural increase was more readily apparent. Examining RMSE under the same permutations showed a reasonably anticipated control for this behavior, but showed similar trends nonetheless:

Table 14.

Simulation Outcomes for Test Length (N_i) and Sample Size (N) Conditions, Continued

Condition	Simplified Distance		RMSE (*b*)		RMSE (*a*)	
	M	(SD)	M	(SD)	M	(SD)
$N_i = 15$						
$N = 50$	7.30	(3.302)	1.33	(1.051)	0.93	(0.909)
$N = 250$	3.68	(0.819)	0.94	(0.707)	0.67	(0.559)
$N = 500$	3.30	(0.393)	0.91	(0.684)	0.64	(0.509)
$N = 1{,}000$	3.12	(0.258)	0.90	(0.676)	0.62	(0.489)
$N_i = 30$						
$N = 50$	10.91	(4.759)	1.35	(1.080)	0.96	(0.940)
$N = 250$	4.98	(0.740)	0.87	(0.637)	0.63	(0.512)
$N = 500$	4.67	(0.404)	0.84	(0.626)	0.62	(0.503)
$N = 1{,}000$	4.30	(0.353)	0.83	(0.623)	0.60	(0.475)

DIF Measurement

One of the straightforward but important purposes of this study was to examine if there was a correlation between the easy-to-calculate Mahalanobis distance based statistic for full models and the DTF statistic when evaluated between reference and focal models. These DIF statistics performed as follows:

Table 15.

Simulation DIF Statistics, by Condition

Condition	DTF		MD	
	M	(SD)	M	(SD)
Test Length				
15	0.68	(1.026)	2.19	(0.433)
30	1.93	(2.703)	3.64	(0.787)
Sample Size				
50	3.31	(4.518)	2.67	(0.863)
250	1.55	(2.177)	2.81	(0.967)
500	1.11	(1.832)	2.92	(0.965)
1000	0.96	(1.585)	3.16	(0.965)
DIF Presence				
(None)	0.52	(1.404)	2.91	(0.966)
20% @ -0.5	0.57	(1.154)	2.92	(0.969)
20% @ -1	1.02	(1.596)	2.94	(0.972)
33% @ -0.5	1.34	(1.975)	2.85	(0.944)
33% @ -1	3.83	(2.694)	2.90	(0.959)
Generating Model				
2PL	1.68	(2.640)	3.04	(0.926)
2PL+C	1.10	(1.885)	2.80	(0.984)
3PL	1.07	(1.619)	2.89	(0.960)

Correlation of DIF Measurements

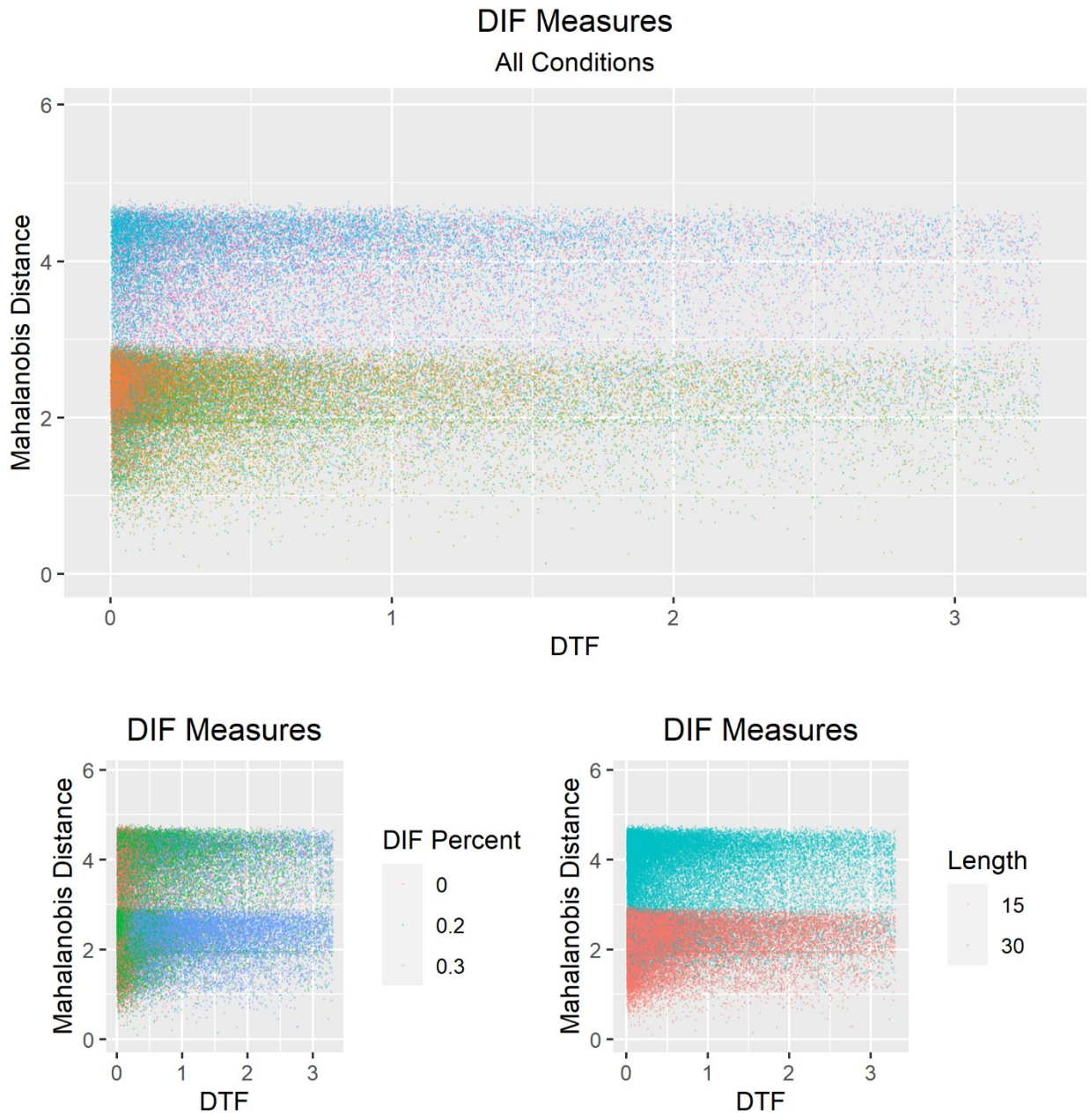
The per-condition examination of correlation between the traditional DTF and new Mahalanobis distance-based DIF measurement statistics is listed in Table 8. The correlations varied: *r* ranged between -.28 and .37 with a median value of 0. The completeness of these variables was also diverse: *df* ranged between 32 and 623. All cases where *df* was below 520 were observed in permutations with the low sample size (N=50) condition; all of the low sample size conditions were below this value. Furthermore, the correlation with a moderate relationship demonstrated only had a *df* of 43 and one of the highest RMSE values for the difficulty parameter; the poor performance of this permutation—another within the small sample size condition—warrants cautionary interpretation. The remaining correlations suggesting a small relationship ($.09 < R^2 > .01$) also occur within the small sample size condition and presented

RMSE on the difficulty parameter higher than the median value. In other words, as the quality of DTF measurements increased, the correlation between them decreased.

Consequentially, analysis of the relationship between these two statistics, across all conditions, showed little meaningful correlation. The new statistic was, probably due to large degrees of freedom, positively correlated to and weakly predicted by DTF, $r = .21$, $F(1, 68922) = 2877$, $p < .01$. Test length remains the most distinguished feature in the relationship, but separating the dataset by test length resulted in contrary correlations, $r(35130) = -.05$, $p < .01$ for 15-item conditions and $r(33790) < .01$, $p < .01$ for 30-item conditions. Some of these relationships were more easily given context when examined visually, such as the aforementioned relationship of interest between established DTF and the new Mahalanobis distance-based statistics.

Figure 10

Measure Comparisons

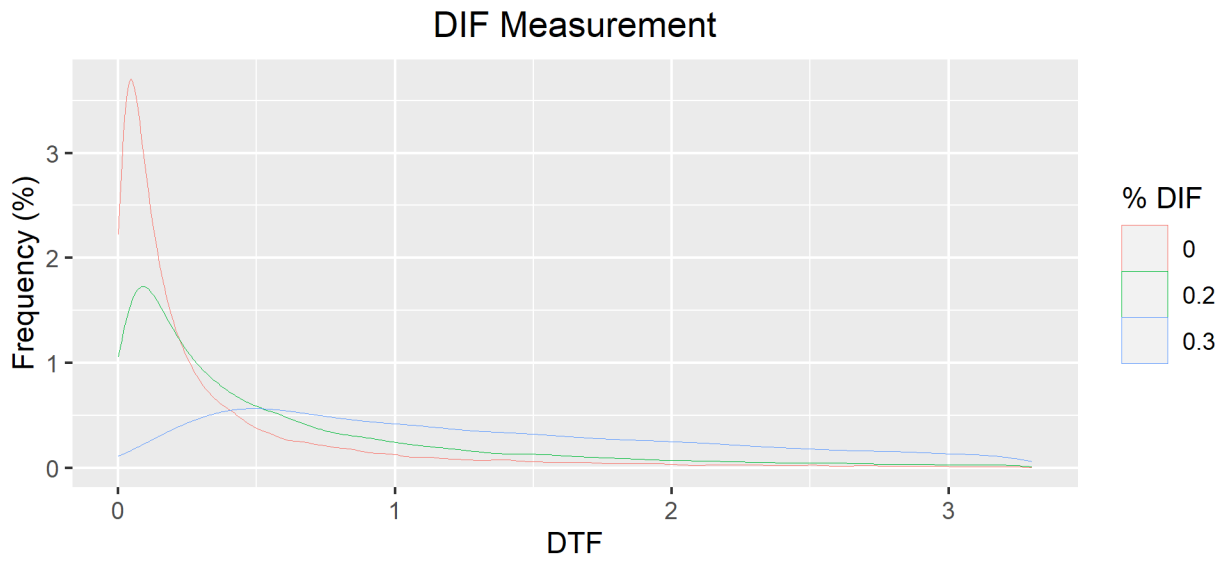


DTF

DTF was measured with the *DFIT* R package. It performed as anticipated with a notable ability to distinguish between models with and without DIF items. Figure 11 depicts the relationship between DTF and DIF and implies a certain amount of empirical power:

Figure 11

DTF across DIF conditions



Specifically, 26% of trials with DIF present had DTF statistics exceeding the 95th percentile of DTF for trials with no DIF present (approximately 2.05). Or, alternatively speaking, 91% of all trials having DTF values greater than that cut-off actually contained DIF. This was with just one iteration of DFIT analysis and randomly-selected anchor items into the bargain. Further graphical examinations of DTF across other conditions are included in Figure 17 in Appendix B. The proportion of DIF trials exceeding that empirical cut-off value for various conditions was:

Table 16

DIF Trial Proportions by Condition

Condition	DTF		P(DIF)
	M	(SD)	
Test Length			
15	0.68	(1.026)	.08
30	1.93	(2.703)	.24
Sample Size			
50	3.31	(4.518)	.14
250	1.55	(2.177)	.21
500	1.11	(1.832)	.16
1000	0.96	(1.585)	.13
DIF Presence			
(None)	0.52	(1.404)	—
20% @ -0.5	0.57	(1.154)	.05
20% @ -1	1.02	(1.596)	.13
33% @ -0.5	1.34	(1.975)	.17
33% @ -1	3.83	(2.694)	.68
Generating Model			
2PL	1.68	(2.640)	.23
2PL+C	1.10	(1.885)	.12
3PL	1.07	(1.619)	.13

Note. P(DIF) = Proportion of DIF trials with DTF > 95th percentile of DTF for all DIF-Free trials

Examining the amount of DIF identified in this manner helps illustrate the improvement of the DTF statistic under certain conditions, most obviously the percent of DIF but also notable is the magnitude of DIF. These proportions should be interpreted with caution in light of the disproportionate amount of errors encountered in evaluating the DTF statistics in the first place. This relationship between DTF and the presence of DIF was analyzed in more detail with the inclusion of other simulation conditions as potentially conflating factors:

Table 17.

Analysis of Variance for DTF and Simulation Conditions

Condition	Sum of Squares	*df*	Mean Square	*F*	*p(F)*	η^2
Test Length	27,022	1	27,022	16,029.30	< .01	x.09
Sample Size	21,221	3	7,074	4,196.00	< .01	x.07
DIF Presence	96,482	4	24,120	14,308.16	< .01	x.31
Generating Model	4,355	2	2,178	1,291.83	< .01	x.01
Length x Size	8,198	3	2,733	1,621.09	< .01	x.03
Length x DIF	22,559	4	5,640	3,345.41	< .01	x.07
Length x Model	1,467	2	734	435.15	< .01	< .01
Size x DIF	1,387	12	116	68.54	< .01	< .01
Size x Model	626	6	104	61.91	< .01	< .01
DIF x Model	3,664	8	458	217.71	< .01	.01
Length x Size x DIF	2,818	12	235	139.28	< .01	.01
Length x Size x Model	82	6	14	8.08	< .01	< .01
Length x DIF x Model	1,652	8	206	122.47	< .01	< .01
Size x DIF x Model	3,355	24	140	82.94	< .01	.01

Using traditional η^2 cut-off values of .01, .06, and .14, multiple influences on DTF can be observed in the study. A large effect was observed from the DIF presence condition; medium effects were observed from the test length and sample size conditions as well as interactions between test length and DIF presence; and small effects were observed from the model type condition as well as interaction effects between test length and sample size as well as DIF presence and generating model. A three-way interaction between sample size, DIF presence, and generating model was observed. Many of these interactions effects align with common sense and were thus desirable as confirmation of the measure working as intended: in so many ways, DTF varied with the varying presence of DIF. Most potential interactions, however, were not found to be practically significant.

Mahalanobis Distance Statistic

The Mahalanobis distance-based statistic did not demonstrate similar amounts of empirical power for the detection of DIF items in a test. It was calculated very quickly using R functions and properties of the *ltm* object. Parameter estimates' standard errors are carefully

extracted into a matrix object from the *ltm* object with the *summary* function to invoke IRT parameterization, then transformed into Mahalanobis distances with the *Mahalanobis* function using the *cov* function on the resulting matrix to transform the coordinates appropriately. Finally, the skewness of the resulting vector can be evaluated via the *psych* library, such that:

$$D = \sqrt{(x-m)^T S^{-1}(x-m)}$$
$$MD = \tilde{\mu}_3 = \frac{\sum_i^{N_i}(D_i-\bar{D})^3}{(N_i-1)\sigma^3} \quad (34)$$

where x is the matrix of row observations of multiple column variables, m is the vector of those variables' means, S is the covariance matrix of those variables, and the *T* denotes a transposed vector, creating a univariate result of appropriately scaled values. This statistic was recorded for each condition as one of the primary outcomes in the study. For example, in trial 75,626:

Table 18.

Example Preparation of Mahalanobis Distance-based Statistic for Trial 75,626

Item	Estimate		Standard Error		MD
	B	A	(B)	(A)	
1	-0.75	0.64	(0.135)	(0.082)	1.64
2	-0.32	1.06	(0.075)	(0.099)	0.75
3	0.29	1.09	(0.077)	(0.097)	0.84
4	-0.71	0.67	(0.126)	(0.084)	1.30
5	-0.26	1.90	(0.053)	(0.152)	3.75
6	1.47	0.52	(0.242)	(0.078)	14.64
7	-0.39	1.09	(0.075)	(0.101)	0.62
8	-0.41	0.80	(0.095)	(0.087)	1.10
9	-1.55	1.15	(0.144)	(0.130)	5.30
10	-0.41	1.37	(0.065)	(0.118)	0.52
11	-0.3	0.96	(0.081)	(0.094)	0.96
12	-0.86	1.06	(0.095)	(0.106)	0.01
13	-0.43	1.89	(0.055)	(0.156)	4.52
14	-0.85	1.12	(0.090)	(0.110)	0.01
15	-0.29	1.74	(0.056)	(0.140)	2.10
16	-0.72	1.23	(0.078)	(0.114)	0.15
17	-0.28	1.66	(0.057)	(0.136)	1.62
18	-0.11	0.83	(0.088)	(0.087)	1.42
19	-0.27	1.60	(0.058)	(0.130)	1.16
20	-0.33	0.83	(0.091)	(0.088)	1.17
21	0.09	0.88	(0.085)	(0.088)	1.41
22	-1.47	0.81	(0.171)	(0.101)	4.55
23	-1.16	1.03	(0.117)	(0.110)	0.56
24	-0.17	0.68	(0.103)	(0.081)	1.60
25	-0.99	1.53	(0.080)	(0.145)	2.91
26	-0.67	1.13	(0.082)	(0.107)	0.13
27	0.41	1.09	(0.079)	(0.097)	0.76
28	-0.81	1.27	(0.081)	(0.119)	0.25
29	-0.54	1.45	(0.065)	(0.126)	0.75
30	-0.04	0.80	(0.091)	(0.085)	1.51

Note. B = Difficulty, A= Discrimination

For this trial, then, the new Mahalanobis Distance-based statistic is the skewness of the Mahalanobis distances of the parameter estimate standard errors, or 3.62. This process was conducted for each full model. Typical values for this statistic between various conditions were:

Table 19

Mahalanobis Distance Statistic Across Conditions, M(SD)

Condition	MD		Time[1]	
	M	(SD)	M	(SD)
Test Length				
15	2.19	(0.433)	0.54	(2.048)
30	3.64	(0.787)	0.55	(0.252)
Sample Size				
50	2.67	(0.863)	0.54	(0.232)
250	2.81	(0.967)	0.54	(0.251)
500	2.92	(0.965)	0.54	(0.223)
1000	3.16	(0.965)	0.56	(2.786)
% DIF Items				
0	2.91	(0.966)	0.54	(0.213)
20%	2.93	(0.970)	0.54	(0.147)
33%	2.88	(0.952)	0.55	(2.532)
Magnitude				
-0.5	2.90	(0.961)	0.55	(2.045)
-1	2.92	(0.965)	0.54	(0.295)
Generating Model				
2PL	3.04	(0.926)	0.55	(2.544)
2PL+C	2.80	(0.984)	0.54	(0.265)
3PL	2.89	(0.960)	0.54	(0.264)

Note. MD = Mahalanobis Distance-based statistic, Time = milliseconds to calculate

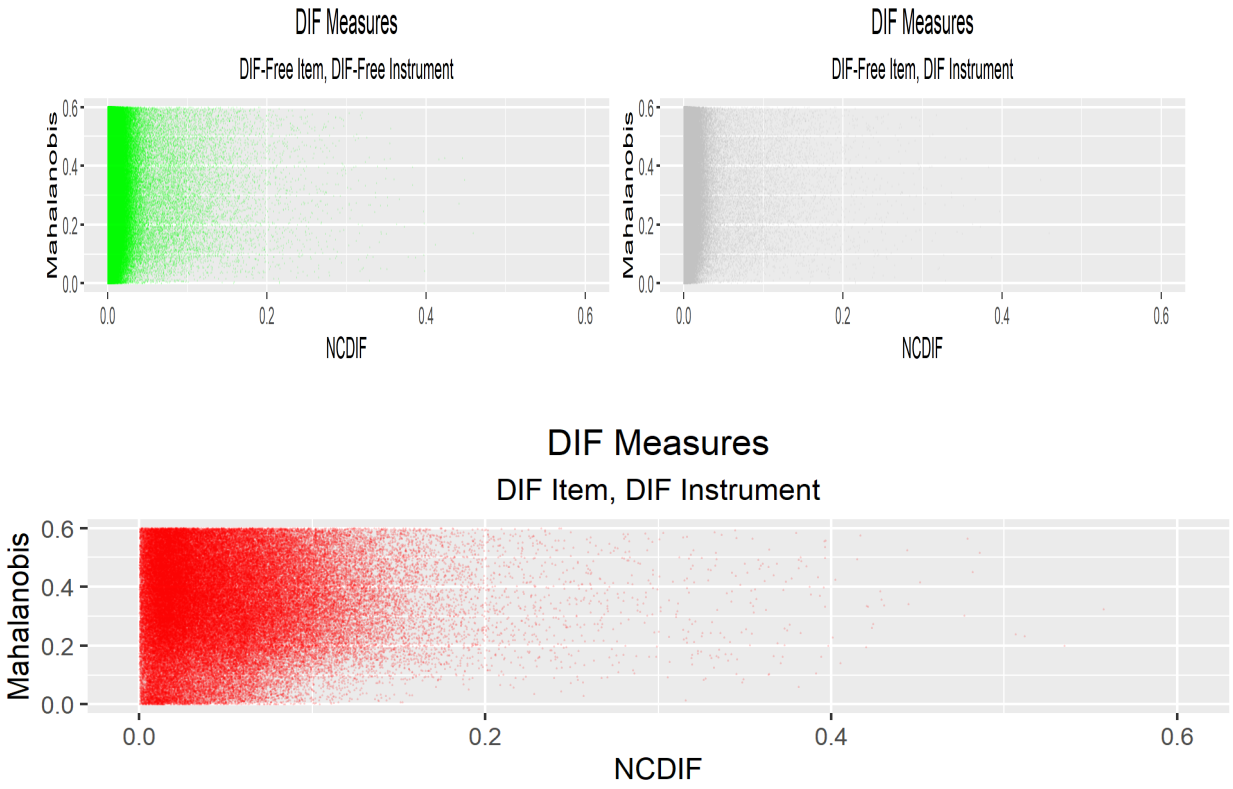
The Mahalanobis distance statistic is also depicted visually throughout various conditions in Appendix B, where it is easier to visually observe differences between conditions. The statistic displayed insensitivity to DIF, DIF magnitude, and underlying model type. The statistic showed some differences, visually, with test length, and only the smallest sample size condition showed different behavior; this is where the smallest variance was observed in the new Mahalanobis distance based statistic, and it was the only condition where values between levels were not within a standard deviation of other values. In other words, the statistic did not vary with the presence of DIF nor the manipulation of simulation conditions other than test length.

Further Item-Level DIF Statistic Analysis

After weak relationships at the instrument level were observed, a deeper examination of item-level performance of the Mahalanobis Distance based Statistic compared to NCDIF was prepared to determine if any correlation was present when instruments were disaggregated. Figure 12 depicts this examination, and illustrates the behavior of the Mahalanobis distance and NCDIF statistic in items with and without DIF; NCDIF displays a more useful, responsive performance. Even at the time level, there was no meaningful correlation between these measurements, $r(1540738) = -.04$, $p < .001$.

Figure 12

NCDIF and new statistic based on item context

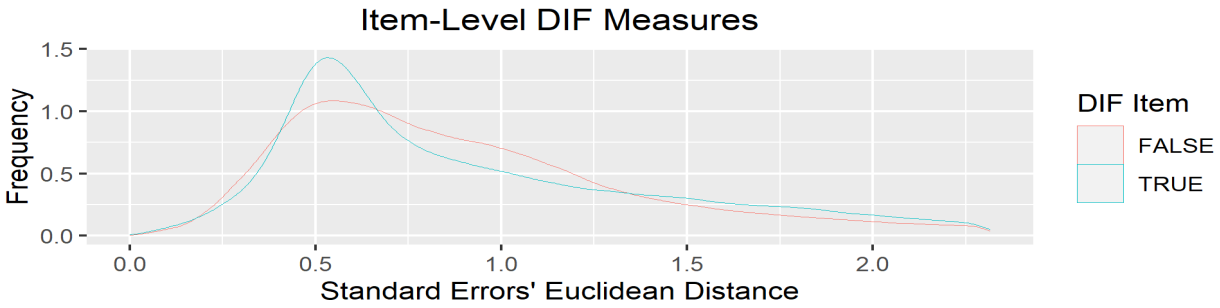


The underlying hypothesis of the study was that standard errors of estimates in a full model may contain useful information on the presence of DIF. To thoroughly explore whether

that hypothesis held true, the basic Euclidean distance used to calculate the rotated and scaled Mahalanobis distance for each item's parameter's standard error was prepared:

Figure 13

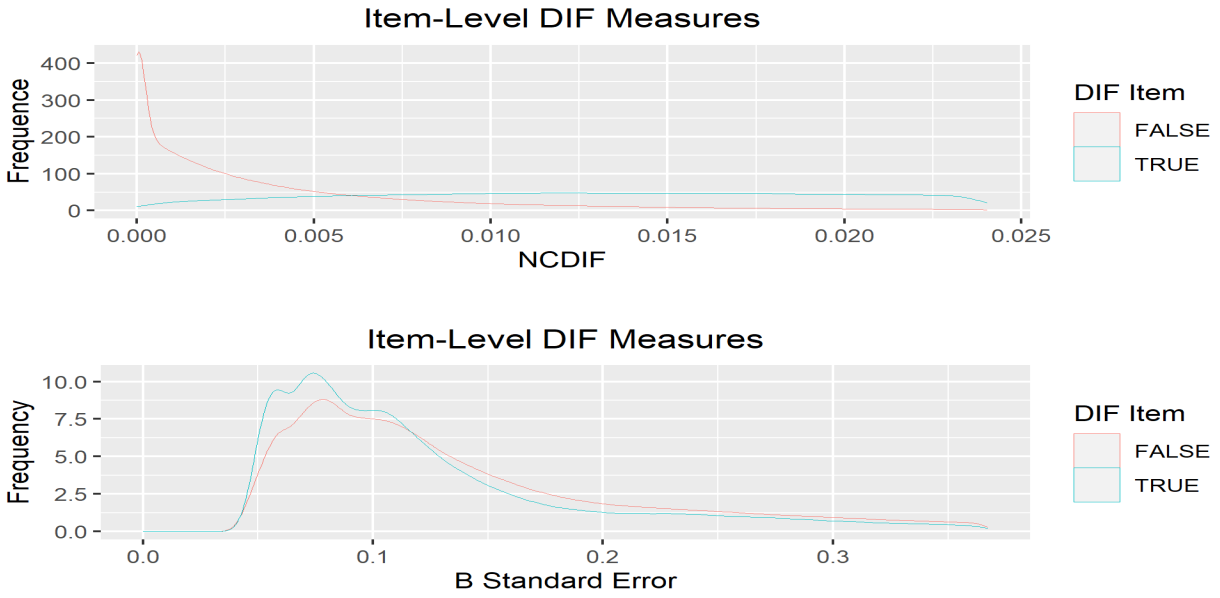
NCDIF and unidimensional item-level standard error of estimates



The results did not demonstrate as clear a relationship between these standard errors and DIF as the NCDIF statistic, which is related to DTF at the item-level, provided. A final comparison was prepared of all 2,023,080 estimated b-parameters and their item's NCDIF statistic:

Figure 14

Item-level performance of NCDIF and b-parameter standard error



The limitation in the hypothesis is clear: parameter standard errors did not strongly react to DIF.

Instruments resilient to DIF items. Since DIF was only injected into the difficulty parameter in the study, a further analysis of variance was conducted using just the performance of difficulty parameter estimates to focus on the influence of DIF rather than an overall model performance with DIF potentially in it:

Table 20.

Analysis of Variance for Difficulty Parameter RMSE by Condition

Condition	Sum of Squares	*df*	Mean Square	*F*	*p(F)*	η^2
Test Length	8	1	8	21.03	< .01	< .01
Sample Size	1,282	3	427	1,177.62	< .01	.04
DIF Presence	22	4	6	15.26	< .01	< .01
Generating Model	434	2	217	598.11	< .01	.01
Length x Size	14	3	5	13.02	< .01	< .01
Length x DIF	17	4	4	11.94	< .01	< .01
Length x Model	7	2	3	9.15	< .01	< .01
Size x DIF	10	12	1	2.34	< .01	< .01
Size x Model	85	6	14	39.06	< .01	< .01
DIF x Model	11	8	1	3.64	< .01	< .01
Length x Size x DIF	60	12	5	13.86	< .01	< .01
Length x Size x Model	23	6	4	10.55	< .01	< .01
Length x DIF x Model	29	8	4	10.10	< .01	< .01
Size x DIF x Model	46	24	2	5.32	< .01	< .01

Based on this analysis, sample size remains a significant factor on the parameter estimation accuracy of models, and the amount of DIF in the study may not have been impactful enough for a DIF item to distort other parameters estimated at the same time. Interestingly, test length no longer appears practically significant from the strict perspective of test-wide RMSE of the difficulty parameter. This should be cautiously interpreted in the context of convergence failures.

Computation Time

Rudimentary timestamp-based methods recorded the duration of each step. This consisted of temporary variables recording the time prior to and just after each calculation. While more robust methods for profiling the performance of R code execution exist, the consistency of this method and its low overhead proved sufficient to serve as a means of comparing conditions and

statistics. Complexity, as might be expected, lead to noticeable increases in computation time. Iterative methods such as Item Parameter Replication would involve even more execution time than the iterations in the study, confirming the problem of scaling time cost. The Mahalanobis Distance-based statistic was evaluated orders of magnitude faster than DTF. A curious, decrease in DTF evaluation time is observed in N=250 conditions; no explanation was found. To wit:

Table 21.

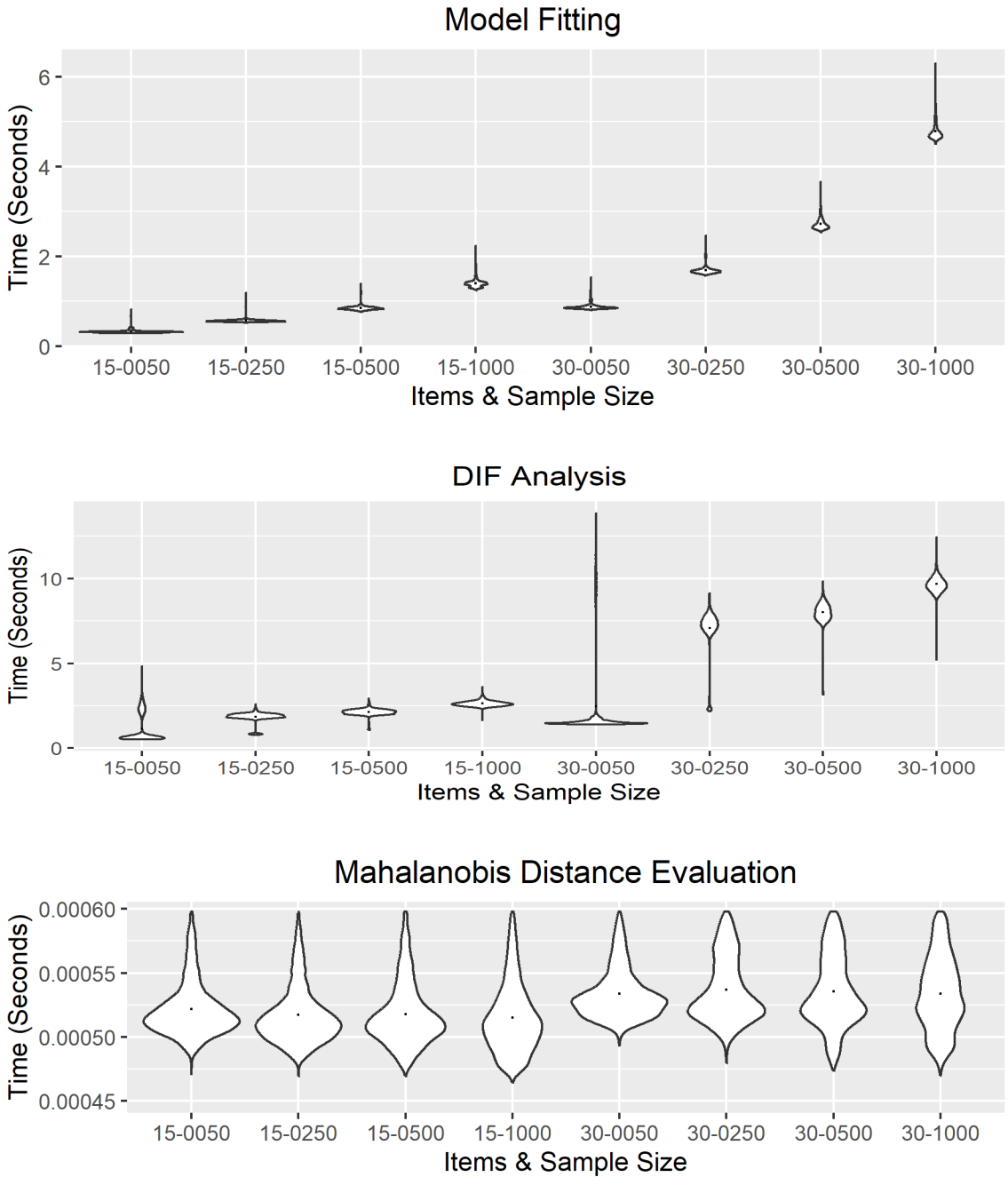
Computation Time, in Seconds, for Test Length (N_i) and Sample Size (N) Conditions

Condition	Full Model		DTF		MD[1]	
	M	(SD)	M	(SD)	M	(SD)
$N_i = 15$						
$N = 50$	0.32	(0.032)	2.41	(0.337)	0.53	(0.260)
$N = 250$	0.57	(0.045)	1.92	(0.128)	0.53	(0.301)
$N = 500$	0.85	(0.057)	2.14	(0.139)	0.53	(0.159)
$N = 1{,}000$	1.40	(0.089)	2.63	(0.149)	0.56	(3.927)
$N_i = 30$						
$N = 50$	0.88	(0.063)	9.74	(1.118)	0.54	(0.197)
$N = 250$	1.69	(0.087)	7.38	(0.433)	0.55	(0.189)
$N = 500$	2.73	(0.127)	8.04	(0.444)	0.55	(0.272)
$N = 1{,}000$	4.74	(0.182)	9.67	(0.423)	0.55	(0.311)

[1] Mahalanobis Distance-based Statistic, time in milliseconds

Figure 15

Computation times for simulation components



5 DISCUSSION

The purpose of this study was to examine the performance of a new statistic, based on implementation of the Mahalanobis distance on the standard errors of model estimators, compared to the DTF statistic. It proposed that current trends in classroom size, scaling costs of DIF analysis, and looming constraints in computational resources create a landscape where the best DIF analyses are becoming more challenging to do well. It theorized that IRT models with DIF might demonstrate detectable patterns in parameter estimation error that could allow for detection of DIF without the need to specify groups nor fit additional models.

Comparison of DIF Measurements

The relationship between the existing DTF statistic and the new Mahalanobis distance-based statistic was hypothesized to justify using the latter as a simplified pre-screening method to efficiently judge when the former may warrant robust evaluation. This relationship would depend on a strong correlation between the statistics throughout the testing conditions in which DIF occurs. The study manipulated some key variables including sample size, test length, and the presence of DIF to simulate these while measuring both DTF and the new Mahalanobis distance-based statistic. Throughout all 120 conditions, however, fewer than 10 displayed a small correlation between these statistics, those correlations themselves lacked statistical significance, however. Only a single condition displayed a moderate correlation, with statistical significance, and none demonstrated a large, significant correlation

The condition with a significant and moderate correlation involved the shorter test, the smallest sample size, the 2PLM with fixed pseudo-random parameter in the generating model, and no DIF presence. Out of all 120 conditions, it presented the sixth highest RMSE on the difficulty parameter, suggesting one of the least reliably interpretable estimates by IRT.

Additionally, this condition set was in the bottom 10% of groups in terms of available degrees of freedom for the correlation evaluation; just as the IRT model performed poorer in this model, so too did the DTF statistic fail to evaluate at an elevated rate. These failures were, coincidentally enough, mostly due to parameter estimates for the focal group failing to converge or, in many cases, producing wild, untenable numbers at convergence unsuitable for DTF calculation.

While these traits could have been interpreted as a relationship that was not resilient in the presence of DIF, the relatively poor performance of IRT and DIF analysis in this condition suggest a different interpretation. As conditions improve to produce more meaningful IRT estimates and DIF measurement, the relationship between DTF and the new Mahalanobis distance-based statistic disappears. This was more likely to be a demonstration of no correlation than of a lost one, then, since the correlation arises out of inaccuracies. This does make a certain amount of sense: since the new Mahalanobis distance-based method is focused on the parameter estimation standard errors of the full model, it could be correlated to a DIF measurement when that DIF measurement only has measurement error to detect. In other words, this could be a further indicator that the new approach is not sensitive to the presence of DIF.

Outcome Interpretations

The study compared the performance of two DIF measurement methods, DTF and a new Mahalanobis distance-based statistic, within the context of a simulation study altering conditions expected to influence parameter estimation effectiveness. Computation time as recorded throughout the study as another dimension for comparison. Several relationships were examined in the course of investigating the way that DIF interacted with parameter estimation, the actual detection and measurement of that DIF, and the computational cost of these IRT models.

Parameter Estimation

Test Length Conditions. Test length had a complex influence on parameter estimation. Although some of the best-estimated models included the longer condition, the proportion remains similar to those observed in the worst models. The variation, however, of the parameter differences varies much more for the smaller condition. In other words, more variety was observed within the N_i=15 condition than between the N_i=15 and N_i=30 conditions. The disproportionate problems with smaller sample sizes were further exacerbated by smaller test length, which may have skewed results.

Sample Size Conditions. The sample size condition was associated with multiple simulation problems, as expected, but in multiple ways that illustrate the perils of using IRT with small datasets. A sample size of 50 is below the recommended size for IRT in the literature (Sahin & Anil, 2017; Uyar & Ozturk Gubes, 2020). Many educational leaders in the United States would preside over schools with grade level memberships of similar size (*Common Core of Data (CCD)*, 2021). As a result, this study specifically included the small sample size condition of N=50 to examine just *how* badly models performed and, potentially, if the new Mahalanobis distance based statistic offered a way to deal with those failures.

Those conditions performed terribly in IRT, and the Mahalanobis distance based statistic did not rescue them. The small sample size condition failed to converge more often than any other condition. Additionally, the models that did converge within the small sample size condition presented more outliers—more bad estimates—than any other condition as well, vastly underperforming every other permutation of the simulation study. Furthermore, within trials where all models converged, DTF statistics failed to evaluate over 10% of the time just in these

smaller sample size conditions. Thus, the new Mahalanobis distance-based statistic did not demonstrate any particular ability to "recover" from these problems small sample sizes present.

Other Conditions. The study altered other conditions including the magnitude of DIF, the proportion of DIF items, and the underlying IRT model selected to generate the data. Within the study, the levels for these conditions did not present practically significant variation to the estimation of parameters when examined at the test level. Possible reasons for this are similar to those in other areas of the study; the sample sizes and test lengths, relative to the smaller proportions of DIF itself, presented enough measurable content that these other fluctuations were not problematic. This aligns with findings that have inspired test-level examinations of DIF with interests in non-compensatory or other practical qualifiers to DIF (Cervantes, 2012; Chalmers et al., 2016; De Boeck & Cho, 2021; Penny, 1994; Wright, 2011). Additionally, one of the condition permutations with the greatest potential amount of DIF with the least potential opportunity to measure it ($N=50$, $N_i=30$, 2PL, 33% DIF @ -1) experienced non-convergence problems for the DTF statistic in the majority of cases.

Summary. Parameter estimation performed as expected in areas where IRT is commonly known to struggle, but some interesting effects were demonstrated by this study. Not every model performed entirely as might be expected by over-application of those common knowledge constraints, however, and the particular ways in which smaller conditions survived IRT analysis may be worthy of further study.

DIF Measurement

Overall, the study found that the DTF statistic can detect the presence of DIF in instruments even with single passes and simple anchoring. The power of DTF is known to increase with the use of such procedures, but the rising computational costs and demands on

sampling data are similarly elevated (Fikis & Oshima, 2017; González-Betanzos & Abad, 2012; Kopf et al., 2014b; Wang, 2004; Woods, 2009; Yuan et al., 2021). This study focused on a quick implementation of DTF that a casual researcher unfamiliar with the subtleties of IRT may apply to a model in the hopes of gaining some assurance that the test in question is fair. The DTF statistic showed an ability to do that, and comparison with the new Mahalanobis distance-based statistic did not demonstrate a similar ability in the proposed, new method. Some variability in the effectiveness of DTF was also observed throughout study conditions.

Focal Group Size. When examining the performance of DTF measurement in the context of information available from the recording of simulation errors, the influence of focal group size as a subset of sample size is apparent. Focal group size was not altered in any simulation conditions; it remained a fixed, small percentage (25%) of sample size representative of some small, under-represented group of interest. It is difficult to interpret these ancillary, implied findings related to focal group size, and other studies have not made a point of directly examining them. The work of Wright (2011), for example, used equal sized focal and reference groups but, like this study, did not vary group size as a condition. Seybert and Stark (2012) was the source for item parameters for the study, but that simulation did not alter its focal group size, either, in the process of item purification. Although this study was able to report how many trials failed to produce DTF statistics and that many of those situations involved small focal group representation, more study may be warranted to produce a better-informed interpretation.

Visual Depictions Suggest Complex Properties. When examining performance visually, such as with the charts in Appendix B, the proportion of DIF items through both test length and DIF percentage is more visible than other conditions. The magnitude of DIF did not present anywhere near the same amount of influence over the model, perhaps being the least

influential condition in the study at every turn. Alterations to the pseudo-random parameter used to generate forms did not appear to interfere with the evaluation of DIF, but that underlying model type presented enough variation to be somewhat noticed in terms of parameter estimation accuracy when examined graphically.

Mahalanobis Distance-based Statistic. The new Mahalanobis Distance-based statistic showed, more than anything else, sensitivity to test length above and beyond what may have been expected from a standardized measurement. It could be that this statistic demonstrates a greater sensitivity to model complexity than to model quality. Multiple DIF detection techniques, in fact, make use of log-odds and other measurements to examine both complexity and fitness to make judgments about DIF in a test (Diaz et al., 2021; Fidalgo & Madeira, 2008; Finch, 2005; Holland & Thayer, 1988). While the statistic did not display sensitivity to DIF within full models without group specification, this observed sensitivity to test item length could imply a suitability for use in more traditional algorithms such as item purification, where simplified models are used to reject bad items on the basis of their failure to improve measurement rather than on the basis of their significant differences in measurement (Fikis & Oshima, 2017; González-Betanzos & Abad, 2012).

Computation Time

The study was able to successfully measure the time required to determine DTF, and that time was observed to scale with model complexity even in one-pass DIF analysis. Although the seconds may at first seem trivial, these costs only compound with each iteration of group definition, and iterative procedures such as item purification and the item parameter replication procedure for evaluating NCDIF only further exacerbate this challenge with unknown costs:

stepwise methods could quickly become prohibitive to the everyday researcher without the resources both in computational time and interpretive time to make meaning from data.

The principle virtue of the Mahalanobis distance-based statistic was that it could be calculated very, very fast—increases measured by factors in the hundreds—and resulted in errors far less often. Compared to the DTF statistic, the resources required scaled much better with the increasing complexity of underlying IRT models throughout the simulation. The Mahalanobis distance-based statistic may not have immediately predicted the presence of DIF, but it had a predictable cost capable of justifying real-time implementation.

Although the Mahalanobis distance did not prove to be a good way to conduct group-free DIF analysis nor provide a "check engine light" for IRT models, the processes of the study did demonstrate the benefits of high-performance computing. There is a chance that mere random sampling could depict more of a relationship between the Mahalanobis distance and DIF than really exists if restricted to the condition sizes and repetition counts allowed with R on typical computers. Indeed, in small trial versions of the study, significant findings were found; only by using the large number of repetitions recommended in the literature was the full picture made apparent, and only through the privilege of supercomputing were those days of computing time made practically possible. With them, the statistics were allowed to demonstrate their true properties, which proved enlightening even if some proved to be uninspiring. This study would have been impossible to conduct without supercomputers; over 5,000 core hours were used in the process of conducting, refining, and analyzing the study. Those resources would not have been accessible without the use of multithreaded approaches to R. Appendix C goes into some detail with portions of R code and commentary on their function and development potential.

Future simulation studies without such luxuries would benefit from considered preparation for dividing the study into more approachable, combinable batches. For example, the hundreds of gigabytes of RAM used for this study were only necessary because of a combination of multi-threading, scaling costs for R and the code structure maintaining all simulation results in the workspace. Other software, such as SAS, uses flat files and routinely moves data from RAM to disk; there are R packages that attempt to replicate this behavior, and in some environments a researcher may find them rewarding.

Implications

The study compared the performance of DTF and a new Mahalanobis distance-based statistic for the pre-screening of IRT models for DIF. Generally, it found that DTF was effective but slow and sensitive, while the new statistic was ineffective but fast and resilient. Conditions in the study demonstrated that DTF was more effective with more DIF to detect in items and persons while some interactions between these aspects could be seen. Parameter estimates, in the presence of DIF, did not always vary. These findings all culminate in implications for both the researcher and policymaker.

Sample Size sets Scope

The main challenge highlighted at the outset of the study was the problem presented by small sample sizes. In typical classrooms and grade levels at some typical schools, there simply would not be enough instruments to conduct a proper IRT analysis, much less a DIF inquiry. The study provided empirical evidence and presented findings of the multi-faceted threats presented by these small sample sizes: existing literature, most probably due to challenges in catching errors in code, did not describe the situation as fully. Both convergence problems and outliers within those remaining, convergent models were the source of most of the messy, “junk” data in

the simulation. Any method that relies on small sample size IRT models, including focal models for DIF tests, takes on this risk; the remaining need for DIF methods that do not specify separately-estimated focal models is highlighted by the implications of this study.

IRT remains Costly

This study helped show how IRT and DIF investigations would not work well in many environments where it could be tempting to use readily available software to point and click into a results table. The implications are bleak: a proper IRT-based DIF analysis remains out of reach for a great proportion of schools without the organizational capacity to implement instrumentation capable of being aggregated into sample sizes reaching into the hundreds. Communicating the rigorous needs of IRT models, particularly the increased demands on models where fairness and DIF analysis are concerns, may be a challenge for educational leaders. Consider, for example, a high stakes test in a small school district where a community is only casually informed about testing. An administrator may have to explain that well-substantiated statements about the fairness of a test depend on the privilege of sample sizes beyond their ability to procure, and this situation may itself have consequences related to the implementation of those high stakes tests in the first place. In other words, the risks of testing might, in these small environments, outweigh the benefits of the measurements produced. Alternatively, these findings might be interpreted to justify collaboration to combine data sources and analytical resources to produce the sample sizes necessary. Some variables remain to be examined, such as focal group size, but this study does provide a substantiated argument for taking steps to amalgamate data, combine groups, and increase sample sizes in ways that might present additional costs for assessment. One simply cannot cut corners when it comes to sample size. Furthermore—even in such cases where “just enough” instruments could be administered to fit a

model—the size of the focal group could itself be small enough to make DIF analysis problematic. The findings in this study show that a DIF analysis can break down at multiple stages, and perhaps the most unlucky cases are those that can and do produce results that are, nevertheless, perilous to interpret.

DIF remains Iterative

DIF measurements in this study did show that simple implementation with even randomly-selected anchors and small focal group specifications can result in successful, reliable detection of DIF. Through, in particular, the significant variation of test length on both DIF measures investigated, iterative procedures remain the most potentially viable way to conduct DIF analysis. Further suggestions for these investigations will follow.

Limitations and Future Research

Linking and Grouping Accuracy

Improper anchor item selection has a negative effect on the functioning of DIF analyses (Fikis & Oshima, 2017). Due to constraints on simulation length, however, conditions including improper anchor item specification could not be conducted. Additionally, the improper identification of group membership has not received much attention in the existing literature. This presents two, related opportunities for further research. First. conditions which intentionally misclassified respondents may produce interesting results; Cappelli (2021) examined the impact of ignoring cross-classification, a more nuanced but related topic, with some improvements observed in that more accurate classification approach. Second, conditions which intentionally select inappropriate or weak anchor items could determine how resilient iterative methods are to inappropriate starting points.

Through alterations to sample size alone, group membership errors could theoretically affect the effectiveness of DIF analyses. With group transience and dynamic identity being a popular topic in modern society, examining the interactions between improperly specified or even less-cohesive groups could shed light on where DIF methods may pass or fail as ways to examine fairness in contexts that are observed with rising frequency in educational settings

Estimated Model Types

Due to constraints on simulation complexity and the unexplored relationships of how the new Mahalanobis distance-based statistic may perform under different dimensionalities, the study could not include conditions where different model types were used to estimate the simulated data. The 2PL was used to evaluate all cases. Conditions with varying models for data generation allowed for some examination of improper model specification, and these conditions provided far less variation in performance of all measures than other conditions. Nevertheless, a potential remains to examine explicitly how estimation with pseudo-random parameters might affect other components observed in the study; Cuhadar et al. (2021) similarly found that ignoring the pseudo-random parameter presented challenges, but the more detailed relationships between parameter omission and the impacts on difficulty parameters observed in that study were not replicated among the results of this study. Future research may benefit from conditions and parameters more similar to that research to generate findings better suited to direct comparison. Han (2012) suggested fixing the pseudo-random parameter similar to the 2PL-C condition used in the study's data generation phase, but that hypothesis was not directly tested with an analysis condition. Another potential improvement or extension of this study would include that now common-place practice to explore its effectiveness and impact on DIF analysis, especially in scenarios where the guessing parameter might differ between groups: c-parameter

based DIF is relatively unexplored in the literature and mathematically hamstrung by common tools used to calculate it. With test preparation curricula including the elimination of "distractor" options as a test-taking strategy, this c-DIF could be the source of no small amount of disparity between groups with varying access to educational resources and a potentially significant area for direct, scholarly investigation.

Iterative Methods

In this study, two DIF measurement devices were compared. First, the established DTF statistic was applied once to a full model, and this simple invocation showed potential usefulness even without iterative refinements. The Item Parameter Replication method allows for determining item-level NCDIF cut-offs, and advances in effect size measurements have been explored in applications of DFIT and DTF mostly related to the Mantel-Haenszel statistic (Finch & French, 2023; Wright & Oshima, 2015).

This study focused on test-wide statistics and, as a result, did not engage in the iterative process of Item Parameter Replication for the determination of cut-off values for individual item parameters as might be conducted in a traditional DFIT analysis. Oshima and Morris (2008) go into some detail on how this process can provide more accurate results by generating NCDIF cut-off values tailored to the properties—namely, the covariance between parameters—of a given instrument. Cervantes (2012) further develops and demonstrates how these covariates being estimated for both focal and reference groups can further improve purification, but this presents even greater mathematical cost and the potential for failed calculations in a DIF analysis. Thus, generally speaking, it is not difficult to find some iterative aspect in most DIF approaches, whether explicitly stepwise or more subtly Bayesian in nature; even likelihood ratio tests involve the progressive comparison of models (Brown et al., 2015). The new Mahalanobis

distance-based statistic did display sensitivity to test length more than any other factor, not unlike the χ^2 statistic Mahalanobis distances are traditionally tested with. Alas, in this study, attempts to invoke χ^2 tests for the full model DIF detection proved unfruitful, though. In spite of this, the Mahalanobis distance based statistic might have potential use in iterative methods such as the full purification method proposed by Fikis and Oshima (2017). In that approach, a test of *i*-items is compared with all possible permutations of *i-1* items for the reduction of evaluated DIF, but it still specifies reference and focal groups; something on the order of *2(i-1)* models must be fit at each step to engage in that process. If the Mahalanobis distance based statistic's sensitivity to test length and interaction between that condition and amount of DIF are truly prominent, as this study suggests, then using it for a similar full purification process could still massively simplify the mathematics involved by avoiding the necessity of group specification and reference and focal group model fitting. With the appropriate multithreaded environment, such feedback could still be within "realtime" perceptibility given the tremendous speed at which this study demonstrated the statistic can be calculated across models of increasing complexity.

A future study examining this property could, for example, prepare permutations of *i-1* instruments and compare how the Mahalanobis distance differences perform relative to typical model fit indices such as the Bayes Information Criterion readily available in the *ltm* package. Examinations of model fit, after all, can use fit indices to demonstrate modelling benefits relative to modelling complexity (Glas & Falcón, 2003; Liang & Wells, 2009; Orlando & Thissen, 2000, 2003).

Geometry

During data cleaning procedures and for some analyses of variance, simplified Euclidean distances of amalgamated item parameter estimation differences—somewhat similar to RMSE—

were used to reduce the dimensionality of a model to something suitable for analysis and reporting. Although the Mahalanobis distance used in creating the new statistic appropriately deals with potential multivariate correlation by scaling and rotating the sample space before determining distances, these basic Euclidean distances did not. The averaged errors in parameters were treated as additive rather than orthogonal on the basis of the central limit theorem, and they were able to produce analysis and findings which conformed to commonly-held wisdom. However, a more rigorous investigation of these relationships and a proof of the relationship between various parameters in an IRT model may be warranted prior to future research. Such an inquiry may even yield insights into further applications of the Mahalanobis distance for detecting DIF in iterative, model fit-based approaches.

Optimizable Code

The simple timestamping used provided clear demonstrations of the rising cost of model fitting and the initial stages of DIF analysis throughout various conditions. There are methods that are more robust; in particular, optimization of memory management is an area for fruitful potential for any researcher engaging with R. Both the *rbenchmark* and *microbenchmark* packages supply functions that could apply in this study, the latter particularly well-suited for relatively fast processes such as the calculation of the Mahalanobis distance. Neither package has, from what can be seen in the literature, been extensively examined in psychometric manuscripts; comparing the two may even be a viable subject for inquiry, much as various model-fitting methods are compared to each other in other studies. Varying methods of measurement could be a worthwhile condition for study, but such research might warrant its own framework. A suitable framework might even go so far as to compare estimation times from other applications such as IRTPRO, adding to the literature in terms of computational cost

where, traditionally, only computational accuracy has been examined (Glas & Falcón, 2003; Kim & Cohen, 1998; Woods et al., 2013).

Cost Assessment. Another dimension of inquiry completely unexamined by this study is financial cost. In risk management, cryptographic methods are expressed in terms of the bottom-line cost in terms of hardware to determine a password; rather than treat a method as foolproof, the decision instead is made on the financial investment of bypassing protection compared to the financial value of whatever is being protected. As this study is particularly interested in the access to IRT in the field, the lack of dollars associated with findings is a gap in its contribution to the literature. In fact, the total financial cost of instrumentation and analysis may be worth preparing; the process as much as the findings could offer real, practical application to educators and administrators both in public and private sectors. A limiting factor is the lifespan of hardware usually used in such calculations: the time required is divided by the lifespan of, for example, a GPU, and the proportion applied to its market price to produce a rough estimate. Naturally, such calculations would not remain current for very long. A researcher interested in this inquiry would be well-advised to prepare formulae that could be invoked by future readers.

Computer Lab Clusters. This study demonstrates the value added from supercomputing environments, but that value is not as privileged as the high sample sizes for classrooms aspiring to IRT analysis! An aspiring researcher could implement SLURM on the workstations in a computer lab or classroom; the idle cycles and vacant RAM of these devices go wasted if not used, and a manuscript on implementing the process would be of interest to at least one researcher. Advances in workspace management, including the availability of natively-supported BASH on Windows, have created an environment ripe with untapped and unprecedented potential. The 200 gigabytes of RAM and 48 CPU cores used in the initial study could easily be

allocated across 20-30 lackluster, idle PCs in a computer lab, and the 18.5 days of core time could turn into one night for such machines. The researcher who develops a tool for a college department to leverage a computer lab into a cluster is one who could save that department no small amount of capital.

Conclusions

The attempt to create a "check engine light" for IRT models did not meet with success in this study. The quest for DIF analysis without group membership specification did not find a panacea with the use of Mahalanobis distances of standard errors of parameter estimates in fitted IRT models in this study, but the quest does not end with it, either. The challenges faced by IRT in terms of scaling time and computational cost, as well as the problematic restrictions of small sample sizes for DIF analysis, are well-established by this study with empirical data. Unfortunately, in spite of empirical data confirming the more calculable nature of the Mahalanobis distance based statistic, the ability to use it to detect DIF in a groupless, full IRT model was not observed. Further investigation may reveal techniques that can make use of the statistic without invoking counter-productive algorithms that undermine its speedy calculation. That speed is the foundation of "realtime," DIF analysis free of anchor item and group specification challenges, and such a method could even empower more automated solutions to account for test fairness during their unsupervised calculations. Such an application warrants further attempts to develop.

This study demonstrated the capabilities of R in high-performance computing contexts, but also suggests some refinement on where to develop potential measurements of DIF and overall IRT model quality. For example, this study has shown that DTF experiences some challenges with low percentages of DIF in smaller sample sizes beyond just the ultimately

problematic N=50 condition; methods to increase focal group size or minimize the impact of that small size such as propensity score matching or the use of correlates may broaden opportunities for DIF analysis that are otherwise demonstrably difficult from this study's findings.

REFERENCES

Ake-Little, E., von der Embse, N., & Dawson, D. (2020). Does Class Size Matter in the University Setting? *Educational Researcher*, *49*(8), 595-605. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1272561&site=ehost-live&scope=site&custid=gsu1 http://dx.doi.org/10.3102/0013189X20933836

Al-zboon, H. S., Alnasraween, M. e. S., & Alkursheh, T. O. (2021). The Effect of the Percentage of Missing Data on Estimating the Standard Error of the Items' Parameters and the Test Information Function According to the Three-Parameter Logistic Model in the Item Response Theory [Article]. *Ilkogretim Online*, *20*(1), 887-898. https://doi.org/10.17051/ilkonline.2021.01.82

An, X., & Yung, Y.-F. (2014). *Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It*. SAS Institute Inc. https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Asianet-Pakistan. (2020). Solon pushes retail price cap on computers. In: Asianet-Pakistan.

Asianet-Pakistan. (2021). Controversial valuation ruling: Price of used computers, laptops jump up to 80pc. In: Asianet-Pakistan.

Bartholomew, D. J., & Shing On, L. (2002). A goodness of fit test for sparse 2p contingency tables [Article]. *British Journal of Mathematical & Statistical Psychology*, *55*(1), 1. https://doi.org/10.1348/000711002159617

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model [https://doi.org/10.1002/j.2333-8504.1981.tb01255.x]. *ETS Research*

Report Series, *1981*(1), i-8. https://doi.org/https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Baudrillard, J. (1994). Simulacra and simulation. http://books.google.com/books?id=EpLXAAAAMAAJ

Bechger, T., & Maris, G. (2015). A Statistical Test for Differential Item Pair Functioning [Article]. *PSYCHOMETRIKA*, *80*(2), 317-340. https://doi.org/10.1007/s11336-014-9408-y

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test [Article]. *Language Testing*, *27*(1), 101-118. https://doi.org/10.1177/0265532209340194

Bialo, J. (2021). *Using Differential Item Functioning and Anchoring Vignettes to Examine the Fairness of Achievement Motivation Items* [Dissertation, Georgia State university]. Atlanta, Georgia.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *PSYCHOMETRIKA*, *46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bohr, M. T. (2018, Mar 13-16). Logic Technology Scaling to Continue Moore's Law. [2018 ieee 2nd electron devices technology and manufacturing conference (edtm 2018)]. IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, JAPAN.

Bowne, J. B., Magnuson, K. A., Schindler, H. S., Duncan, G. J., & Yoshikawa, H. (2017). A Meta-Analysis of Class Sizes and Ratios in Early Childhood Education Programs: Are Thresholds of Quality Associated with Greater Impacts on Cognitive, Achievement, and Socioemotional Outcomes? *Educational Evaluation and Policy Analysis*, *39*(3), 407-428.

https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1149537&site=ehost-live&scope=site&custid=gsu1

http://journals.sagepub.com/doi/full/10.3102/0162373716689489

Brown, C., Templin, J., & Cohen, A. (2015). Comparing the Two- and Three-Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem [Article]. *Applied Psychological Measurement*, *39*(5), 335-348. https://doi.org/10.1177/0146621614563326

Burkman, A., Garrett, J., & Posner, B. Z. (2019). Does Organizational Size Impact the Leadership Practices of School Leaders? *International Journal of Educational Leadership Preparation*, *14*(1), 13-21. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1218872&site=ehost-live&scope=site&custid=gsu1

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm [Article]. *British Journal of Mathematical & Statistical Psychology*, *61*(2), 309-329. https://doi.org/10.1348/000711007X249603

Canbeldek, M., & Isikoglu Erdogan, N. (2017). The Effects of Early Childhood Classroom Size and Duration on Development of Children. *Eurasian Journal of Educational Research*(68), 257-271. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1148856&site=ehost-live&scope=site&custid=gsu1

Cappelli, C. (2021). *Individual mobility across clusters: The impact of ignoring cross-classified data structures in discrete-time survival analysis* Georgia State University]. Atlanta, Georgia.

Cervantes, V. H. (2012). On using the Item Parameter Replication (IPR) approach for power calculation of the noncompensatory differential itemfunctioning (NCDIF) index. Proceedings of the V European Congress of Methodology,

Cervantes, V. H. (2017). DFIT: An R Package for Raju's Differential Functioning of Items and Tests Framework. *Journal of Statistical Software*, *76*(5), 1-24. https://doi.org/10.18637/jss.v076.i05

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement*, *76*(1), 114-140. https://doi.org/10.1177/0013164415584576

Common Core of Data (CCD). (2021).

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

CRAN Task View: Psychometric Models and Methods. (2022). https://cran.r-project.org/web/views/Psychometrics.html

Crane, P. K., Belle, G. v., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI [https://doi.org/10.1002/sim.1713]. *Statistics in Medicine*, *23*(2), 241-256. https://doi.org/https://doi.org/10.1002/sim.1713

Crocker, L. M. A. J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cuhadar, I., Yang, Y., & Paek, I. (2021). Consequences of Ignoring Guessing Effects on Measurement Invariance Analysis. *Applied Psychological Measurement*, *45*(4), 283-296. https://doi.org/10.1177/01466216211013915

Culligan, B. (2011). *Item Response Theory, Reliability and Standard Error*. Aoyama Gakuin Women's Junior College.

Culpepper, S. A. (2016). Revisiting the 4-Parameter Item Response Model: Bayesian Estimation and Application [Article]. *PSYCHOMETRIKA*, *81*(4), 1142-1163. https://doi.org/10.1007/s11336-015-9477-6

Darrell Bock, R., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *PSYCHOMETRIKA*, *35*(2), 179-197. https://doi.org/10.1007/BF02291262

De Boeck, P., & Cho, S. J. (2021). Not all DIF is shaped similarly. *PSYCHOMETRIKA*, *86*(3), 712-716. https://doi.org/10.1007/s11336-021-09772-3

De Paola, M., Ponzo, M., & Scoppa, V. (2013). Class size effects on student achievement: heterogeneity across abilities and fields [Article]. *Education Economics*, *21*(2), 135-153. https://doi.org/10.1080/09645292.2010.511811

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38. http://www.jstor.org/stable/2984875

Diaz, E., Brooks, G., & Johanson, G. (2021). Detecting Differential Item Functioning: Item Response Theory Methods versus the Mantel-Haenszel Procedure. *International Journal of Assessment Tools in Education*, *8*(2), 376-393. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1303867&site=ehost-live&scope=site&custid=gsu1

Dimitrov, D. M. (2017). Examining Differential Item Functioning: IRT-Based Detection in the Framework of Confirmatory Factor Analysis. *Measurement and Evaluation in*

Counseling and Development, *50*(3), 183-200. https://doi.org/10.1080/07481756.2017.1320946

Dorans, N. J., & Holland, P. W. (1992). DIF DETECTION AND DESCRIPTION: MANTEL-HAENSZEL AND STANDARDIZATION1,2 [https://doi.org/10.1002/j.2333-8504.1992.tb01440.x]. *ETS Research Report Series*, *1992*(1), i-40. https://doi.org/https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Draxler, C. (2010). Sample Size Determination for Rasch Model Tests [Article]. *PSYCHOMETRIKA*, *75*(4), 708-724. https://doi.org/10.1007/s11336-010-9182-4

Duncan, S. C. (2006). *Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel -Haenszel DIF methods* (Publication Number 3270730) [Ph.D., Texas A&M University]. ProQuest Dissertations & Theses A&I. Ann Arbor. https://www.proquest.com/dissertations-theses/improving-prediction-differential-item/docview/304934936/se-2?accountid=11226

Elosua, P., & Wells, C. (2013). Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. *Psicologica: International Journal of Methodology and Experimental Psychology*, *34*(2), 327-342. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1019157&site=ehost-live&scope=site&custid=gsu1

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory* [Book]. Psychology Press. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=nlebk&AN=44641&site=ehost-live&scope=site&custid=gsu1

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection. *Educational and Psychological Measurement*, *68*(6), 940-958. https://doi.org/10.1177/0013164408315265

Fidalgo, Á. M., & Scalon, J. D. (2009). Using Generalized Mantel-Haenszel Statistics to Assess DIF Among Multiple Groups. *Journal of Psychoeducational Assessment*, *28*(1), 60-69. https://doi.org/10.1177/0734282909337302

Fidell, L. S., & Tabachnick, B. G. (2003). Preparatory Data Analysis. In *Handbook of Psychology* (pp. 115-141). https://doi.org/https://doi.org/10.1002/0471264385.wei0205

Fikis, D. R. J., & Oshima, T. C. (2017). Effect of Purification Procedures on DIF Analysis in IRTPRO. *Educational and Psychological Measurement*, *77*(3), 415-428. https://doi.org/10.1177/0013164416645844

Finch, H. (2005). The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, *29*(4), 278-295. https://doi.org/10.1177/0146621605275728

Finch, W., & French, B. (2023). Effect Sizes for Estimating Differential Item Functioning Influence at the Test Level. *Psych*, *5*, 133-147. https://doi.org/10.3390/psych5010013

Fischer, G. H. M. I. W. W. (1995). *Rasch Models foundations, recent developments, and applications; [papers originally presented at a workshop held at the University of Vienna, Feb. 25-27, 1993]*. Springer.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A Description and Demonstration of the Polytomous-DFIT Framework. *Applied Psychological Measurement*, *23*(4), 309-326. https://doi.org/10.1177/01466219922031437

GaDOE. (2021). *Georgia Milestones Statewide 2021-2022 Statewide Scores*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia_2021-2022_Assessment_Results.aspx

Garcia, B. (2020). *Coronavirus hits PC hardwire prices in Kuwait* [Article]. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=n5h&AN=2W63255800879&site=ehost-live&scope=site&custid=gsu1

Glas, C. A. W., & Falcón, J. C. S. (2003). A Comparison of Item-Fit Statistics for the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *27*(2), 87-106. https://doi.org/10.1177/0146621602250530

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rash models--foundations, recent developments and applications* (pp. 69-95). Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Tests of fit for polytomous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rash models--foundations, recent developments and applications* (pp. 325-352). Springer.

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(4), 134-145. https://doi.org/10.1027/1614-2241/a000046

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory : principles and applications*. Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2010). *Fundamentals of item response theory* ([Nachdr.]. ed.). Sage.

Han, K. T. (2012). Fixing the c Parameter in the Three-Parameter Logistic Model. *Practical Assessment, Research & Evaluation*, *17*(1). https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ977575&site=ehost-live&scope=site&custid=gsu1

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, *20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an eternal golden braid*. Basic Books.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Erlbaum.

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch Model [Editorial Material]. *Language Assessment Quarterly*, *13*(2), 124-141. https://doi.org/10.1080/15434303.2016.1160096

Ippel, L., & Magis, D. (2020). Efficient Standard Errors in Item Response Theory Models for Short Tests. *Educational and Psychological Measurement*, *80*(3), 461-475. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1253240&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1177/0013164419882072

Isaacson, E., & Keller, H. B. (1966). *Analysis of numerical methods*. Wiley.

Jin-Shei, L., Jeanne, T., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (dif) for small sample sizes [Article]. *Evaluation & the Health Professions*, *28*(3), 283-294. https://doi.org/10.1177/0163278705278276

Kendall, M. G. S. A. (1961). The Advanced theory of statistics. Inference and relationship. Vol. 2 Vol. 2. http://catalog.hathitrust.org/api/volumes/oclc/25351202.html

Kennedy, T. (2006). *Some notes on sufficient statistics*. University of Arizona. https://www.math.arizona.edu/~tgk/466/sufficient.pdf

Kieftenbeld, V., & Natesan, P. (2012). Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*, *36*(5), 399-419. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ970948&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1177/0146621612446170

Kim, K. Y., & Lee, W.-C. (2017). The Impact of Three Factors on the Recovery of Item Parameters for the Three-Parameter Logistic Model. *Applied Measurement in Education*, *30*(3), 228-242. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1144205&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1080/08957347.2017.1316274

Kim, S.-H., & Cohen, A. S. (1998). A Comparison of Linking and Concurrent Calibration Under Item Response Theory. *Applied Psychological Measurement*, *22*(2), 131-143. https://doi.org/10.1177/01466216980222003

Kissell, R., & Poserina, J. (2017). Advanced math and statistics. In *Optimal sports math, statistics, and fantasy*. Academic Press.

Kopf, J., Zeileis, A., & Strobl, C. (2014a). Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educational and Psychological Measurement*, *75*(1), 22-56. https://doi.org/10.1177/0013164414529792

Kopf, J., Zeileis, A., & Strobl, C. (2014b). A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection. *Applied Psychological Measurement*, *39*(2), 83-103. https://doi.org/10.1177/0146621614544195

Laitsch, D., Nguyen, H., & Younghusband, C. H. (2021). Class Size and Teacher Work: Research Provided to the BCTF in Their Struggle to Negotiate Teacher Working Conditions. *Canadian Journal of Educational Administration and Policy*(196), 83-101. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1301589&site=ehost-live&scope=site&custid=gsu1

Langer, M. (2008). *A Reexamination of Lord's Wald Test for Differential Item Functioning Using Item Response Theory and Modern Error Estimation* [Dissertation, University of North Carolina at Chapel Hill]. Carolina Digital Repository.

Leesa-Nguansuk, S. (2021). *Price of new computers on the rise* [Article]. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=n5h&AN=2W61825901435&site=ehost-live&scope=site&custid=gsu1

Leesa-Nguansuk, S. (2022). *Computer shortage looming* [Article]. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=n5h&AN=2W62697918071&site=ehost-live&scope=site&custid=gsu1

Li, W., & Konstantopoulos, S. (2016). Class Size Effects on Fourth-Grade Mathematics Achievement: Evidence from TIMSS 2011. *Journal of Research on Educational Effectiveness*, *9*(4), 503-530. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1115245&site=ehost-live&scope=site&custid=gsu1 http://dx.doi.org/10.1080/19345747.2015.1105893

Li, W., & Konstantopoulos, S. (2017). Does Class-Size Reduction Close the Achievement Gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement*, *28*(2), 292-313. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1139569&site=ehost-live&scope=site&custid=gsu1 http://dx.doi.org/10.1080/09243453.2017.1280062

Liang, T., & Wells, C. S. (2009). A Model Fit Statistic for Generalized Partial Credit Model. *Educational and Psychological Measurement*, *69*(6), 913-928. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ863225&site=ehost-live&scope=site&custid=gsu1 http://dx.doi.org/10.1177/0013164409332222

Lin, C.-H., Kwon, J. B., & Zhang, Y. (2019). Online Self-Paced High-School Class Size and Student Achievement. *Educational Technology Research and Development*, *67*(2), 317-336. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1208091&site=ehost-live&scope=site&custid=gsu1 http://dx.doi.org/10.1007/s11423-018-9614-x

Linden, W. J. v. d. H. R. K. (1996). *Handbook of modern item response theory*. Springer.

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, *7*. http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc., 365 Broadway, Hillsdale, NJ 07642 ($39.95).

Lord, F. M. (1984). Standard Errors of Measurement at Different Ability Levels. *Journal of Educational Measurement*, *21*(3), 239-243. http://www.jstor.org/stable/1434781

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Pub. Co. https://search.amphilsoc.org/collections/view?docId=ead/Mss.Ms.Coll.117-ead.xml;query=tukey;brand=default

Lord, F. M., Tukey, J. W., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Pub. Co. https://search.amphilsoc.org/collections/view?docId=ead/Mss.Ms.Coll.117-ead.xml;query=tukey;brand=default

Lowenthal, P. R., Nyland, R., Jung, E., Dunlap, J. C., & Kepka, J. (2019). Does Class Size Matter? An Exploration into Faculty Perceptions of Teaching High-Enrollment Online Courses. *American Journal of Distance Education*, *33*(3), 152-168. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1220254&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1080/08923647.2019.1610262

M2PressWIRE. (2020). Global GPU Industry Outlook and Forecast to 2027: A $451+ Billion Opportunity. In: M2PressWIRE.

Magis, D., & Facon, B. (2013). Item Purification Does Not Always Improve DIF Detection: A Counterexample With Angoff's Delta Plot. *Educational and Psychological Measurement*, *73*(2), 293-311. https://doi.org/10.1177/0013164412451903

Mahalanobis, P. C. (1936). On the Generalised Distance in Statistics (Reprint, 2018). *Sankhya A*, *80*(1), 1-7. https://doi.org/10.1007/s13171-019-00164-5

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute*, *22*(4), 719-748. https://doi.org/10.1093/jnci/22.4.719

Mazor, K. M., & et al. (1991). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. In.

McBride, J., & Weiss, D. J. (1974). A word knowledge item pool for adaptive ability measurement (ED096339). *ERIC*.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing [Article]. *Language Testing*, *29*(4), 555-576. https://doi.org/10.1177/0265532211430367

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Muraki, E. (1992). A GENERALIZED PARTIAL CREDIT MODEL: APPLICATION OF AN EM ALGORITHM [https://doi.org/10.1002/j.2333-8504.1992.tb01436.x]. *ETS Research Report Series*, *1992*(1), i-30. https://doi.org/https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models / edited by Michael L. Nering, Remo Ostini*. Routledge.

O'Boyle, M. (2019, Feb 16-20). Rethinking Compilation in a Heterogeneous World (Keynote).*International Symposium on Code Generation and Optimization* [Proceedings of the 2019 ieee/acm international symposium on code generation and optimization (cgo '19)]. 17th IEEE/ACM International Symposium on Code Generation and Optimization (CGO), Washington, DC.

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of Sample Size on Common Item Equating Using the Dichotomous Rasch Model. *Applied Measurement in Education*, *33*(1), 10-23. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1244757&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1080/08957347.2019.1674309

Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50-64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *27*(4), 289-298. https://doi.org/10.1177/0146621603027004004

Oshima, T. C., & Morris, S. B. (2008). An NCME Instructional Module on Raju's Differential Functioning of Items and Tests (DFIT) [Article]. *Educational Measurement: Issues & Practice*, *27*(3), 43-50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Oshima, T. C., Raju, N., & Nanda, A. O. (2006). A New Method for Assessing the Statistical Significance in the Differential Functioning of Items and Tests (DFIT) Framework

[Article]. *Journal of Educational Measurement*, *43*(1), 1-17. https://doi.org/10.1111/j.1745-3984.2006.00001.x

Pekmezci, F. B., & Avsar, A. S. (2021). A Guide for More Accurate and Precise Estimations in Simulative Unidimensional IRT Models [Article]. *International Journal of Assessment Tools in Education*, *8*(2), 423-453. https://doi.org/10.21449/ijate.790289

Penny, J. A. (1994). *Using the area between two item response functions to index differential item functioning: a generalized approach* University of North Carolina at Greensboro]. http://libres.uncg.edu/ir/uncg/listing.aspx?styp=ti&id=27172

Pohl, S., Schulze, D., & Stets, E. (2021). Partial Measurement Invariance: Extending and Evaluating the Cluster Approach for Identifying Anchor Items. *Applied Psychological Measurement*, *45*(7-8), 477-493, Article 01466216211042809. https://doi.org/10.1177/01466216211042809

Putz, B., Kam-Thong, T., Karbalai, N., Altmann, A., & Muller-Myhsok, B. (2013). Cost-effective GPU-Grid for Genome-wide Epistasis Calculations [Article]. *Methods of Information in Medicine*, *52*(1), 91-95. https://doi.org/10.3414/me11-02-0049

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

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests [Article]. *Applied Psychological Measurement*, *19*(4), 353-368. https://doi.org/10.1177/014662169501900405

Randall, J., & Engelhard, G., Jr. (2010). Using Confirmatory Factor Analysis and the Rasch Model to Assess Measurement Invariance in a High Stakes Reading Assessment

[Article]. *Applied Measurement in Education*, *23*(3), 286-306. https://doi.org/10.1080/08957347.2010.486289

Rao, C. R. (1965). *Linear statistical inference and its applications*. John Wiley and Sons.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks pædagogiske Institut.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methodse Anthony S. Bryk and Stephen W. Raudenbush* (2nd ed.). Sage Pubns.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLOS ONE*, *11*(4), e0151138. https://doi.org/10.1371/journal.pone.0151138

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133-144. http://www.jstor.org/stable/1434973

Reporter, D. M. (2021). Chip shortage causes price hike [Article]. *Daily Mail*, 14. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=n5h&AN=152108027&site=ehost-live&scope=site&custid=gsu1

Robitzsch, A., & Ludtke, O. (2022). Mean Comparisons of Many Groups in the Presence of DIF: An Evaluation of Linking and Concurrent Scaling Approaches. *JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS*, *47*(1), 36-68, Article 10769986211017479. https://doi.org/10.3102/10769986211017479

Rotman, D. (2020). The end of the greatest prediction on earth [Article]. *MIT Technology Review*, *123*(2), 10-13.

https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=fth&AN=141708713&site=eds-live&scope=site&custid=gsu1

Sahin, A., & Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory and Practice*, *17*, 321-335. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1130806&site=ehost-live&scope=site&custid=gsu1

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores1 [https://doi.org/10.1002/j.2333-8504.1968.tb00153.x]. *ETS Research Bulletin Series*, *1968*(1), i-169. https://doi.org/https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Schulze, D., Reuter, B., & Pohl, S. (2022). Measurement Invariance: Dealing with the Uncertainty in Anchor Item Choice by Model Averaging. *STRUCTURAL EQUATION MODELING-A MULTIDISCIPLINARY JOURNAL*, *29*(4), 521-530. https://doi.org/10.1080/10705511.2021.2012785

Seybert, J., & Stark, S. (2012). Iterative Linking With the Differential Functioning of Items and Tests (DFIT) Method: Comparison of Testwide and Item Parameter Replication (IPR) Critical Values. *Applied Psychological Measurement*, *36*(6), 494-515. https://doi.org/10.1177/0146621612445182

Shen, T., & Konstantopoulos, S. (2021). Estimating Causal Effects of Class Size in Secondary Education: Evidence from TIMSS. *Research Papers in Education*, *36*(5), 507-541. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1310848&site=ehost-live&scope=site&custid=gsu1

http://dx.doi.org/10.1080/02671522.2019.1697733

Sheng, Y. Y., Welling, W. S., & Zhu, M. M. (2014, Jul 21-25). GPU-Accelerated Computing with Gibbs Sampler for the 2PNO IRT Model.*Springer Proceedings in Mathematics & Statistics* [Quantitative psychology research]. 79th Annual Meeting of the Psychometric Society, Univ Wisconsin, Madison, WI.

Sorensen, C. (2015). An Examination of the Relationship between Online Class Size and Instructor Performance. *Journal of Educators Online*, *12*(1), 140-159. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1051032&site=ehost-live&scope=site&custid=gsu1

Stewart, J. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly*, *11*(3), 271-282. https://doi.org/10.1080/15434303.2014.922977

Stroud, A. H., & Secrest, D. (1966). *Gaussian Quadrature Formulas*.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Harper & Row.

Templin, J. (2008). *Framing Item Response Models as Hierarchical Linear Models* Hierarchical Linear Models Workshop, http://www.edmeasurement.net/8268/Templin-2008-IRT-as-HLM.pdf

Uyar, S., & Ozturk Gubes, N. (2020). Item Parameter Estimation for Dichotomous Items Based on Item Response Theory: Comparison of BILOG-MG, Mplus and R (ltm) [Article]. *Journal of Measurement and Evaluation in Education and Psychology-Epod*, *11*(1), 27-42. https://doi.org/10.21031/epod.591415

Wald, A. (1942). Asymptotically Shortest Confidence Intervals. *The Annals of Mathematical Statistics*, *13*(2), 127-137. http://www.jstor.org/stable/2235750

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, *54*(3), 426-482. https://doi.org/10.2307/1990256

Waller, M. I. (1981). A Procedure for Comparing Logistic Latent Trait Models. *Journal of Educational Measurement*, *18*(2), 119-125. http://www.jstor.org/stable/1434653

Wang, W.-c. (2004). Effects of Anchor Item Methods on the Detection of Differential Item Functioning Within the Family of Rasch Models [Article]. *Journal of Experimental Education*, *72*(3), 221-261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, Y. Y., Bowers, A. J., & Fikis, D. J. (2017). Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles From 1965 to 2014 [Review]. *Educational Administration Quarterly*, *53*(2), 289-323. https://doi.org/10.1177/0013161x16660585

Warm, T. A. (1978). *A Primer of Item Response Theory: Technical Report 940279*.

Wen-Chung, W., & Ya-Hui, S. (2004). Factors Influencing the Mantel and Generalized Mantel-Haenszel Methods for the Assessment of Differential Item Functioning in Polytomous Items [Article]. *Applied Psychological Measurement*, *28*(6), 450-480. https://doi.org/10.1177/0146621604269792

Wilks, S. S. (1938). Shortest Average Confidence Intervals from Large Samples. *The Annals of Mathematical Statistics*, *9*(3), 166-175. http://www.jstor.org/stable/2957730

Wilks, S. S. (1944). *Mathematical statistics*. Princeton University Press.

Woods, C. M. (2009). Empirical Selection of Anchors for Tests of Differential Item Functioning [Article]. *Applied Psychological Measurement*, *33*(1), 42-57. https://doi.org/10.1177/0146621607314044

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT [Article]. *Educational & Psychological Measurement*, *73*(3), 532-547. https://doi.org/10.1177/0013164412464875

Wright, B., & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement*, *29*(1), 23-48. https://doi.org/10.1177/001316446902900102

Wright, K. D. (2011). *Improvements for Differential Functioning of Items and Tests (DFIT): Investigating the Addition of Reporting an Effect Size Measure and Power* [Dissertation, Georgia State University]. Atlanta, Georgia.

Wright, K. D., & Oshima, T. C. (2015). An Effect Size Measure for Raju's Differential Functioning for Items and Tests [Article]. *Educational and Psychological Measurement*, *75*(2), 338-358. https://doi.org/10.1177/0013164414532944

Yuan, K. H., Liu, H. Y., & Han, Y. T. (2021). Differential Item Functioning Analysis Without A Priori Information on Anchor Items: QQ Plots and Graphical Test. *PSYCHOMETRIKA*, *86*(2), 345-377. https://doi.org/10.1007/s11336-021-09746-5

Zumbo, B. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*.

APPENDICES

Appendix A: Relative full model parameter estimates across conditions

Table 22

Relative Estimation Differences, B Parameter, by Condition, M(SD)

Item	Sample Size				DIF Percent		
	50	250	500	1000	0	20%	33%
1	-0.68 (1.63)	-0.45 (0.63)	-0.46 (0.6)	-0.43 (0.57)	-0.49 (0.88)	-0.47 (0.87)	-0.51 (0.94)
2	-0.21 (0.62)	-0.17 (0.41)	-0.2 (0.41)	-0.18 (0.4)	-0.09 (0.4)	-0.23 (0.48)	-0.24 (0.47)
3	0.26 (0.59)	0.26 (0.38)	0.24 (0.36)	0.25 (0.36)	0.25 (0.4)	0.26 (0.43)	0.24 (0.42)
4	-0.48 (1.16)	-0.39 (0.55)	-0.41 (0.54)	-0.39 (0.52)	-0.37 (0.7)	-0.35 (0.66)	-0.51 (0.73)
5	-0.2 (0.47)	-0.18 (0.38)	-0.21 (0.38)	-0.19 (0.37)	-0.19 (0.38)	-0.18 (0.39)	-0.21 (0.4)
6	1.32 (3.11)	1.52 (1.5)	1.4 (1.17)	1.37 (1.08)	1.4 (1.78)	1.42 (1.82)	1.41 (1.7)
7	-0.27 (0.63)	-0.24 (0.43)	-0.26 (0.43)	-0.24 (0.41)	-0.21 (0.43)	-0.2 (0.45)	-0.35 (0.51)
8	-0.49 (1.33)	-0.4 (0.56)	-0.42 (0.54)	-0.4 (0.52)	-0.42 (0.71)	-0.41 (0.75)	-0.44 (0.81)
9	-1.27 (1.78)	-1.12 (0.98)	-1.14 (0.98)	-1.11 (0.95)	-1.15 (1.18)	-1.13 (1.14)	-1.18 (1.19)
10	-0.29 (0.58)	-0.27 (0.43)	-0.3 (0.43)	-0.28 (0.42)	-0.19 (0.4)	-0.32 (0.48)	-0.33 (0.48)
11	-0.08 (0.65)	-0.05 (0.38)	-0.07 (0.37)	-0.05 (0.36)	-0.06 (0.41)	-0.05 (0.42)	-0.08 (0.47)
12	-0.68 (1.12)	-0.59 (0.62)	-0.61 (0.62)	-0.58 (0.6)	-0.61 (0.73)	-0.59 (0.73)	-0.63 (0.75)
13	-0.28 (0.53)	-0.27 (0.42)	-0.29 (0.42)	-0.27 (0.41)	-0.24 (0.41)	-0.23 (0.41)	-0.38 (0.49)
14	-0.53 (0.81)	-0.48 (0.54)	-0.5 (0.54)	-0.47 (0.52)	-0.49 (0.57)	-0.48 (0.58)	-0.51 (0.63)
15	-0.18 (0.47)	-0.16 (0.37)	-0.18 (0.37)	-0.16 (0.36)	-0.17 (0.38)	-0.16 (0.39)	-0.18 (0.4)

Table 23

Relative Estimation Differences, B Parameter, by Condition, M(SD) (continued)

Item	Sample Size				DIF Percent		
	50	250	500	1000	0	20%	33%
16	-0.51 (0.88)	-0.44 (0.52)	-0.46 (0.53)	-0.43 (0.5)	-0.46 (0.57)	-0.42 (0.54)	-0.49 (0.67)
17	-0.24 (0.5)	-0.23 (0.4)	-0.25 (0.41)	-0.22 (0.39)	-0.15 (0.37)	-0.26 (0.42)	-0.3 (0.45)
18	-0.12 (0.82)	-0.06 (0.4)	-0.08 (0.4)	-0.05 (0.37)	-0.07 (0.49)	-0.05 (0.48)	-0.1 (0.53)
19	-0.23 (0.52)	-0.2 (0.39)	-0.22 (0.4)	-0.2 (0.38)	-0.17 (0.38)	-0.14 (0.38)	-0.32 (0.46)
20	-0.33 (1.22)	-0.27 (0.51)	-0.3 (0.51)	-0.26 (0.47)	-0.29 (0.66)	-0.26 (0.67)	-0.31 (0.73)
21	0.18 (0.71)	0.19 (0.39)	0.16 (0.37)	0.19 (0.36)	0.18 (0.4)	0.2 (0.44)	0.15 (0.51)
22	-1.26 (2.12)	-1.1 (1)	-1.12 (0.99)	-1.07 (0.94)	-1.1 (1.29)	-1.07 (1.27)	-1.21 (1.24)
23	-1.02 (1.76)	-0.91 (0.84)	-0.93 (0.85)	-0.88 (0.8)	-0.93 (1.09)	-0.9 (0.98)	-0.96 (1.12)
24	-0.06 (0.89)	-0.02 (0.41)	-0.05 (0.4)	-0.02 (0.38)	-0.04 (0.48)	-0.01 (0.53)	-0.06 (0.55)
25	-0.89 (1.15)	-0.83 (0.75)	-0.85 (0.77)	-0.81 (0.73)	-0.77 (0.8)	-0.86 (0.81)	-0.9 (0.9)
26	-0.5 (0.89)	-0.43 (0.53)	-0.46 (0.54)	-0.42 (0.51)	-0.45 (0.59)	-0.42 (0.56)	-0.49 (0.67)
27	0.42 (0.64)	0.42 (0.45)	0.39 (0.42)	0.42 (0.42)	0.42 (0.46)	0.43 (0.5)	0.38 (0.47)
28	-0.69 (0.97)	-0.63 (0.63)	-0.65 (0.65)	-0.62 (0.61)	-0.61 (0.68)	-0.58 (0.64)	-0.74 (0.79)
29	-0.38 (0.61)	-0.34 (0.45)	-0.37 (0.46)	-0.33 (0.44)	-0.35 (0.47)	-0.32 (0.46)	-0.38 (0.52)
30	0.1 (1.03)	0.11 (0.4)	0.09 (0.39)	0.11 (0.37)	0.1 (0.57)	0.13 (0.6)	0.08 (0.52)

Table 24

Relative Estimation Differences, B Parameter, by Condition, M(SD) (continued)

Item	Test Length		Magnitude		Generating Model		
	15	30	-.05	-1	2PL	2PL+C	3PL
1	-0.48 (0.82)	-0.5 (0.97)	-0.51 (0.91)	-0.48 (0.88)	-0.1 (0.49)	-0.71 (1.04)	-0.65 (0.92)
2	-0.18 (0.46)	-0.19 (0.45)	-0.17 (0.45)	-0.21 (0.47)	0.03 (0.32)	-0.35 (0.5)	-0.23 (0.43)
3	0.26 (0.42)	0.24 (0.42)	0.25 (0.42)	0.26 (0.42)	0.39 (0.4)	0.06 (0.4)	0.32 (0.38)
4	-0.41 (0.69)	-0.42 (0.72)	-0.41 (0.71)	-0.41 (0.69)	-0.11 (0.38)	-0.59 (0.79)	-0.52 (0.74)
5	-0.19 (0.39)	-0.2 (0.4)	-0.2 (0.4)	-0.19 (0.39)	-0.03 (0.3)	-0.34 (0.44)	-0.21 (0.37)
6	1.42 (1.7)	1.4 (1.83)	1.42 (1.76)	1.41 (1.78)	1.59 (1.6)	1.18 (1.83)	1.47 (1.84)
7	-0.25 (0.47)	-0.26 (0.47)	-0.25 (0.47)	-0.26 (0.47)	-0.05 (0.33)	-0.42 (0.55)	-0.27 (0.43)
8	-0.42 (0.72)	-0.43 (0.79)	-0.43 (0.77)	-0.41 (0.74)	-0.11 (0.39)	-0.6 (0.9)	-0.54 (0.77)
9	-1.16 (1.15)	-1.15 (1.18)	-1.17 (1.2)	-1.14 (1.13)	-0.98 (0.9)	-1.27 (1.31)	-1.2 (1.23)
10	-0.28 (0.46)	-0.29 (0.46)	-0.26 (0.44)	-0.3 (0.47)	-0.12 (0.34)	-0.44 (0.54)	-0.28 (0.41)
11	-0.06 (0.44)	-0.07 (0.43)	-0.07 (0.44)	-0.06 (0.43)	0.12 (0.32)	-0.29 (0.51)	-0.01 (0.35)
12	-0.61 (0.74)	-0.61 (0.74)	-0.62 (0.76)	-0.6 (0.72)	-0.39 (0.48)	-0.77 (0.88)	-0.66 (0.73)
13	-0.27 (0.44)	-0.29 (0.45)	-0.28 (0.44)	-0.28 (0.45)	-0.13 (0.35)	-0.43 (0.51)	-0.26 (0.4)
14	-0.49 (0.58)	-0.5 (0.6)	-0.5 (0.6)	-0.48 (0.59)	-0.31 (0.42)	-0.65 (0.72)	-0.5 (0.55)
15	-0.16 (0.39)	-0.18 (0.39)	-0.18 (0.39)	-0.16 (0.39)	0 (0.3)	-0.32 (0.43)	-0.18 (0.36)

Table 25

Relative Estimation Differences, B Parameter, by Condition, M(SD) (continued)

Item	Test Length		Magnitude		Generating Model		
	15	30	-.05	-1	2PL	2PL+C	3PL
16	—	-0.46	-0.47	-0.44	-0.24	-0.6	-0.52
	—	(0.6)	(0.63)	(0.57)	(0.39)	(0.69)	(0.6)
17	—	-0.24	-0.22	-0.25	-0.09	-0.39	-0.22
	—	(0.42)	(0.42)	(0.42)	(0.33)	(0.48)	(0.38)
18	—	-0.07	-0.09	-0.06	0.21	-0.23	-0.19
	—	(0.5)	(0.53)	(0.47)	(0.34)	(0.55)	(0.46)
19	—	-0.21	-0.21	-0.21	-0.06	-0.37	-0.2
	—	(0.42)	(0.42)	(0.42)	(0.33)	(0.47)	(0.37)
20	—	-0.29	-0.3	-0.27	0.07	-0.42	-0.5
	—	(0.69)	(0.69)	(0.69)	(0.35)	(0.72)	(0.77)
21	—	0.18	0.17	0.19	0.38	-0.03	0.19
	—	(0.45)	(0.44)	(0.47)	(0.41)	(0.49)	(0.36)
22	—	-1.13	-1.14	-1.11	-0.86	-1.25	-1.25
	—	(1.27)	(1.32)	(1.22)	(0.86)	(1.39)	(1.42)
23	—	-0.93	-0.94	-0.92	-0.68	-1.04	-1.05
	—	(1.07)	(1.06)	(1.08)	(0.67)	(1.2)	(1.18)
24	—	-0.04	-0.05	-0.02	0.26	-0.19	-0.16
	—	(0.52)	(0.53)	(0.51)	(0.41)	(0.54)	(0.49)
25	—	-0.84	-0.84	-0.84	-0.7	-0.94	-0.88
	—	(0.84)	(0.85)	(0.83)	(0.67)	(0.92)	(0.89)
26	—	-0.45	-0.46	-0.44	-0.2	-0.56	-0.59
	—	(0.61)	(0.64)	(0.58)	(0.37)	(0.64)	(0.68)
27	—	0.41	0.4	0.42	0.55	0.23	0.45
	—	(0.48)	(0.46)	(0.49)	(0.49)	(0.43)	(0.45)
28	—	-0.64	-0.65	-0.64	-0.43	-0.75	-0.73
	—	(0.71)	(0.73)	(0.68)	(0.5)	(0.76)	(0.78)
29	—	-0.35	-0.37	-0.34	-0.18	-0.49	-0.38
	—	(0.48)	(0.49)	(0.47)	(0.35)	(0.55)	(0.47)
30	—	0.1	0.09	0.12	0.39	-0.06	0
	—	(0.57)	(0.52)	(0.61)	(0.46)	(0.63)	(0.48)

Table 26

Relative Estimation Differences, A Parameter, by Condition, M(SD)

Item	Sample Size				DIF Percent		
	50	250	500	1000	0	20%	33%
1	-0.01 (0.61)	-0.06 (0.32)	-0.07 (0.28)	-0.07 (0.26)	-0.06 (0.38)	-0.05 (0.37)	-0.07 (0.36)
2	0.6 (1.12)	0.56 (0.65)	0.55 (0.61)	0.55 (0.59)	0.54 (0.76)	0.55 (0.71)	0.6 (0.75)
3	1.08 (1.79)	0.97 (0.97)	0.96 (0.93)	0.96 (0.91)	0.98 (1.16)	1.01 (1.17)	0.96 (1.11)
4	0.17 (0.79)	0.14 (0.42)	0.13 (0.38)	0.13 (0.36)	0.12 (0.48)	0.13 (0.48)	0.16 (0.5)
5	1.81 (2.46)	1.65 (1.36)	1.63 (1.29)	1.62 (1.26)	1.68 (1.63)	1.71 (1.66)	1.6 (1.48)
6	-0.17 (0.72)	-0.2 (0.36)	-0.21 (0.32)	-0.21 (0.3)	-0.2 (0.43)	-0.19 (0.44)	-0.21 (0.42)
7	0.72 (1.3)	0.67 (0.7)	0.66 (0.66)	0.66 (0.64)	0.65 (0.85)	0.67 (0.83)	0.7 (0.8)
8	0.12 (0.76)	0.1 (0.39)	0.09 (0.35)	0.09 (0.33)	0.1 (0.47)	0.11 (0.46)	0.08 (0.46)
9	0.79 (1.87)	0.6 (0.66)	0.58 (0.6)	0.58 (0.57)	0.64 (1.08)	0.65 (1)	0.59 (0.88)
10	1.16 (1.59)	1.09 (0.94)	1.07 (0.9)	1.07 (0.88)	1.08 (1.12)	1.05 (1)	1.14 (1.09)
11	0.49 (1.07)	0.45 (0.58)	0.44 (0.55)	0.45 (0.53)	0.46 (0.68)	0.48 (0.71)	0.43 (0.65)
12	0.46 (1.12)	0.39 (0.54)	0.38 (0.5)	0.39 (0.48)	0.41 (0.67)	0.42 (0.68)	0.38 (0.64)
13	1.62 (2.23)	1.47 (1.18)	1.44 (1.12)	1.44 (1.1)	1.47 (1.5)	1.49 (1.43)	1.49 (1.3)
14	0.84 (1.46)	0.74 (0.73)	0.73 (0.69)	0.73 (0.67)	0.76 (0.89)	0.79 (0.94)	0.72 (0.84)
15	1.44 (2.01)	1.33 (1.12)	1.32 (1.09)	1.32 (1.07)	1.36 (1.41)	1.37 (1.3)	1.3 (1.23)

Table 27

Relative Estimation Differences, A Parameter, by Condition, M(SD) (continued)

Item	Sample Size				DIF Percent		
	50	250	500	1000	0	20%	33%
16	0.66 (1.2)	0.6 (0.63)	0.58 (0.6)	0.58 (0.59)	0.6 (0.79)	0.61 (0.74)	0.58 (0.72)
17	1.51 (1.89)	1.35 (1.08)	1.33 (1.04)	1.33 (1.03)	1.36 (1.33)	1.32 (1.16)	1.43 (1.26)
18	0.3 (0.85)	0.27 (0.48)	0.27 (0.46)	0.27 (0.45)	0.27 (0.56)	0.29 (0.56)	0.26 (0.54)
19	1.51 (2.03)	1.34 (1.07)	1.32 (1.05)	1.32 (1.03)	1.33 (1.38)	1.35 (1.26)	1.4 (1.22)
20	0.11 (0.74)	0.08 (0.38)	0.07 (0.35)	0.07 (0.33)	0.07 (0.44)	0.09 (0.46)	0.07 (0.45)
21	0.43 (1.04)	0.39 (0.56)	0.38 (0.53)	0.38 (0.52)	0.4 (0.7)	0.4 (0.64)	0.38 (0.64)
22	0.22 (1.08)	0.14 (0.41)	0.12 (0.36)	0.13 (0.34)	0.13 (0.53)	0.14 (0.64)	0.16 (0.54)
23	0.4 (1.26)	0.3 (0.48)	0.28 (0.44)	0.28 (0.42)	0.31 (0.7)	0.32 (0.71)	0.3 (0.62)
24	0.23 (0.81)	0.19 (0.45)	0.19 (0.42)	0.19 (0.41)	0.19 (0.51)	0.21 (0.53)	0.19 (0.52)
25	1.14 (1.77)	0.96 (0.84)	0.92 (0.78)	0.93 (0.76)	0.96 (1.2)	0.95 (0.93)	1.01 (0.99)
26	0.62 (1.11)	0.55 (0.62)	0.54 (0.58)	0.54 (0.57)	0.55 (0.72)	0.57 (0.72)	0.54 (0.71)
27	0.93 (1.68)	0.83 (0.9)	0.83 (0.87)	0.82 (0.86)	0.84 (1.13)	0.86 (1.06)	0.83 (1.02)
28	0.77 (1.33)	0.66 (0.67)	0.64 (0.63)	0.64 (0.61)	0.64 (0.83)	0.66 (0.8)	0.71 (0.81)
29	1.23 (1.89)	1.09 (0.92)	1.07 (0.88)	1.06 (0.86)	1.1 (1.16)	1.13 (1.17)	1.08 (1.08)
30	0.18 (0.8)	0.16 (0.44)	0.16 (0.41)	0.16 (0.4)	0.16 (0.51)	0.17 (0.51)	0.16 (0.51)

Table 28

Relative Estimation Differences, A Parameter, by Condition, M(SD) (continued)

Item	Test Length		Magnitude		Generating Model		
	15	30	-.05	-1	2PL	2PL+C	3PL
1	-0.06 (0.38)	-0.06 (0.36)	-0.06 (0.37)	-0.05 (0.37)	0.06 (0.39)	-0.12 (0.35)	-0.1 (0.35)
2	0.56 (0.75)	0.57 (0.73)	0.54 (0.73)	0.58 (0.74)	0.85 (0.81)	0.38 (0.64)	0.47 (0.67)
3	1 (1.16)	0.97 (1.13)	0.98 (1.13)	0.99 (1.17)	1.49 (1.27)	0.42 (0.71)	1.06 (1.13)
4	0.14 (0.5)	0.14 (0.48)	0.13 (0.48)	0.15 (0.49)	0.29 (0.51)	0.06 (0.46)	0.08 (0.46)
5	1.66 (1.56)	1.67 (1.63)	1.67 (1.6)	1.66 (1.58)	2.36 (1.77)	1.17 (1.31)	1.51 (1.43)
6	-0.2 (0.44)	-0.2 (0.42)	-0.2 (0.44)	-0.2 (0.43)	0.08 (0.48)	-0.39 (0.31)	-0.28 (0.33)
7	0.67 (0.82)	0.67 (0.83)	0.65 (0.81)	0.69 (0.84)	0.94 (0.88)	0.48 (0.77)	0.61 (0.76)
8	0.1 (0.48)	0.1 (0.45)	0.09 (0.46)	0.1 (0.47)	0.24 (0.49)	0.01 (0.44)	0.04 (0.43)
9	0.6 (0.91)	0.65 (1.06)	0.62 (1.01)	0.63 (0.97)	0.69 (0.82)	0.58 (1.05)	0.61 (1.07)
10	1.08 (1.08)	1.1 (1.06)	1.08 (1.06)	1.1 (1.08)	1.45 (1.18)	0.81 (0.9)	1.04 (1.01)
11	0.46 (0.69)	0.45 (0.67)	0.45 (0.69)	0.46 (0.67)	0.67 (0.72)	0.23 (0.59)	0.48 (0.65)
12	0.4 (0.67)	0.41 (0.66)	0.4 (0.67)	0.41 (0.66)	0.54 (0.65)	0.31 (0.67)	0.37 (0.65)
13	1.47 (1.39)	1.49 (1.44)	1.47 (1.44)	1.5 (1.39)	1.93 (1.57)	1.08 (1.15)	1.46 (1.37)
14	0.75 (0.89)	0.76 (0.9)	0.75 (0.91)	0.76 (0.87)	0.95 (0.9)	0.6 (0.84)	0.73 (0.9)
15	1.35 (1.32)	1.34 (1.31)	1.34 (1.31)	1.35 (1.32)	1.9 (1.49)	0.94 (1.06)	1.23 (1.19)

Table 29

Relative Estimation Differences, A Parameter, by Condition, M(SD) (continued)

Item	Test Length		Magnitude		Generating Model		
	15	30	-.05	-1	2PL	2PL+C	3PL
16	—	0.6	0.6	0.6	0.79	0.48	0.54
	—	(0.75)	(0.74)	(0.77)	(0.78)	(0.74)	(0.69)
17	—	1.37	1.38	1.35	1.75	0.97	1.41
	—	(1.25)	(1.29)	(1.21)	(1.33)	(0.99)	(1.29)
18	—	0.27	0.28	0.27	0.53	0.13	0.17
	—	(0.56)	(0.56)	(0.55)	(0.63)	(0.47)	(0.46)
19	—	1.36	1.36	1.36	1.77	0.95	1.38
	—	(1.29)	(1.33)	(1.25)	(1.38)	(0.99)	(1.34)
20	—	0.08	0.08	0.07	0.25	0	-0.01
	—	(0.45)	(0.46)	(0.45)	(0.49)	(0.42)	(0.4)
21	—	0.39	0.4	0.39	0.67	0.14	0.38
	—	(0.66)	(0.68)	(0.64)	(0.71)	(0.49)	(0.66)
22	—	0.14	0.14	0.15	0.2	0.12	0.12
	—	(0.57)	(0.58)	(0.56)	(0.55)	(0.64)	(0.51)
23	—	0.31	0.31	0.31	0.38	0.28	0.27
	—	(0.68)	(0.67)	(0.69)	(0.55)	(0.79)	(0.66)
24	—	0.2	0.2	0.2	0.44	0.07	0.11
	—	(0.52)	(0.52)	(0.52)	(0.59)	(0.44)	(0.44)
25	—	0.97	0.96	0.99	1.07	0.9	0.96
	—	(1.05)	(1.06)	(1.04)	(0.95)	(1.04)	(1.14)
26	—	0.56	0.56	0.55	0.79	0.45	0.44
	—	(0.72)	(0.72)	(0.71)	(0.78)	(0.68)	(0.64)
27	—	0.84	0.84	0.84	1.43	0.28	0.85
	—	(1.07)	(1.06)	(1.09)	(1.3)	(0.55)	(0.92)
28	—	0.67	0.66	0.68	0.83	0.58	0.6
	—	(0.81)	(0.82)	(0.8)	(0.84)	(0.8)	(0.78)
29	—	1.1	1.12	1.08	1.42	0.85	1.05
	—	(1.14)	(1.14)	(1.14)	(1.2)	(1.06)	(1.09)
30	—	0.16	0.16	0.16	0.41	0.02	0.07
	—	(0.51)	(0.51)	(0.51)	(0.57)	(0.42)	(0.43)

Appendix B: Full Model DIF Analyses Performance

Figure 16

DTF performance among all conditions

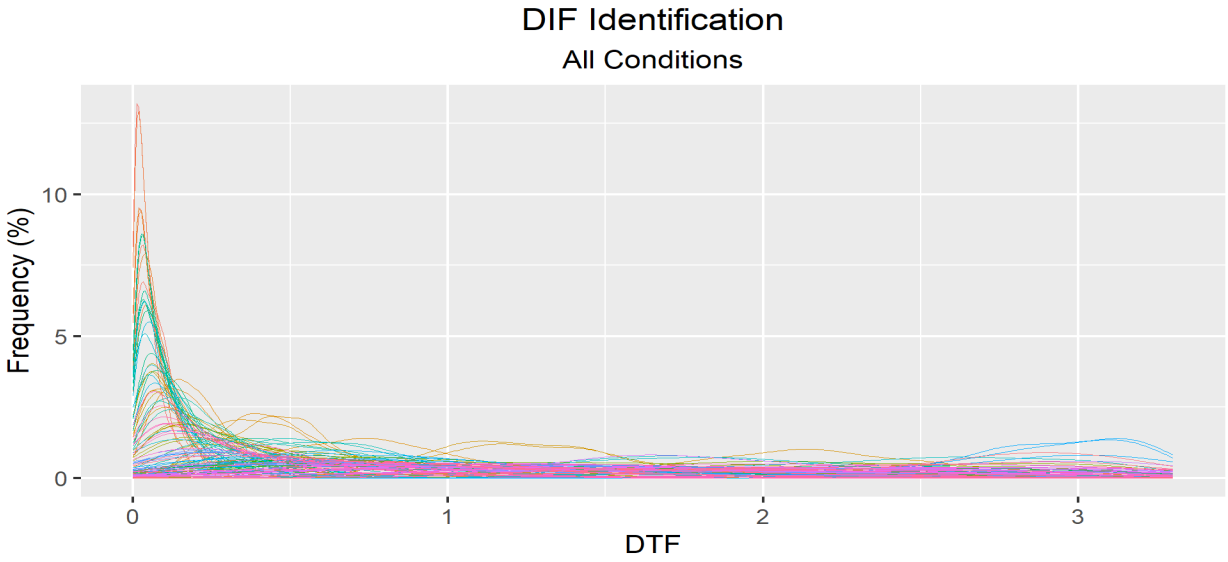


Figure 17

Mahalanobis distance performance among all conditions

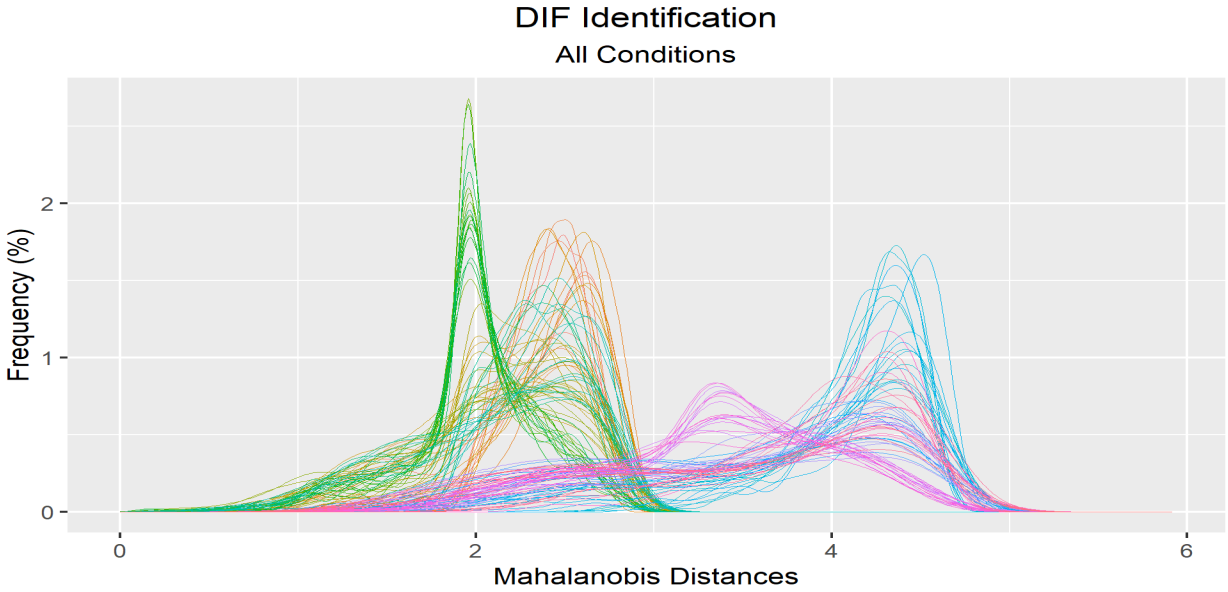


Figure 18

DTF performance within various conditions

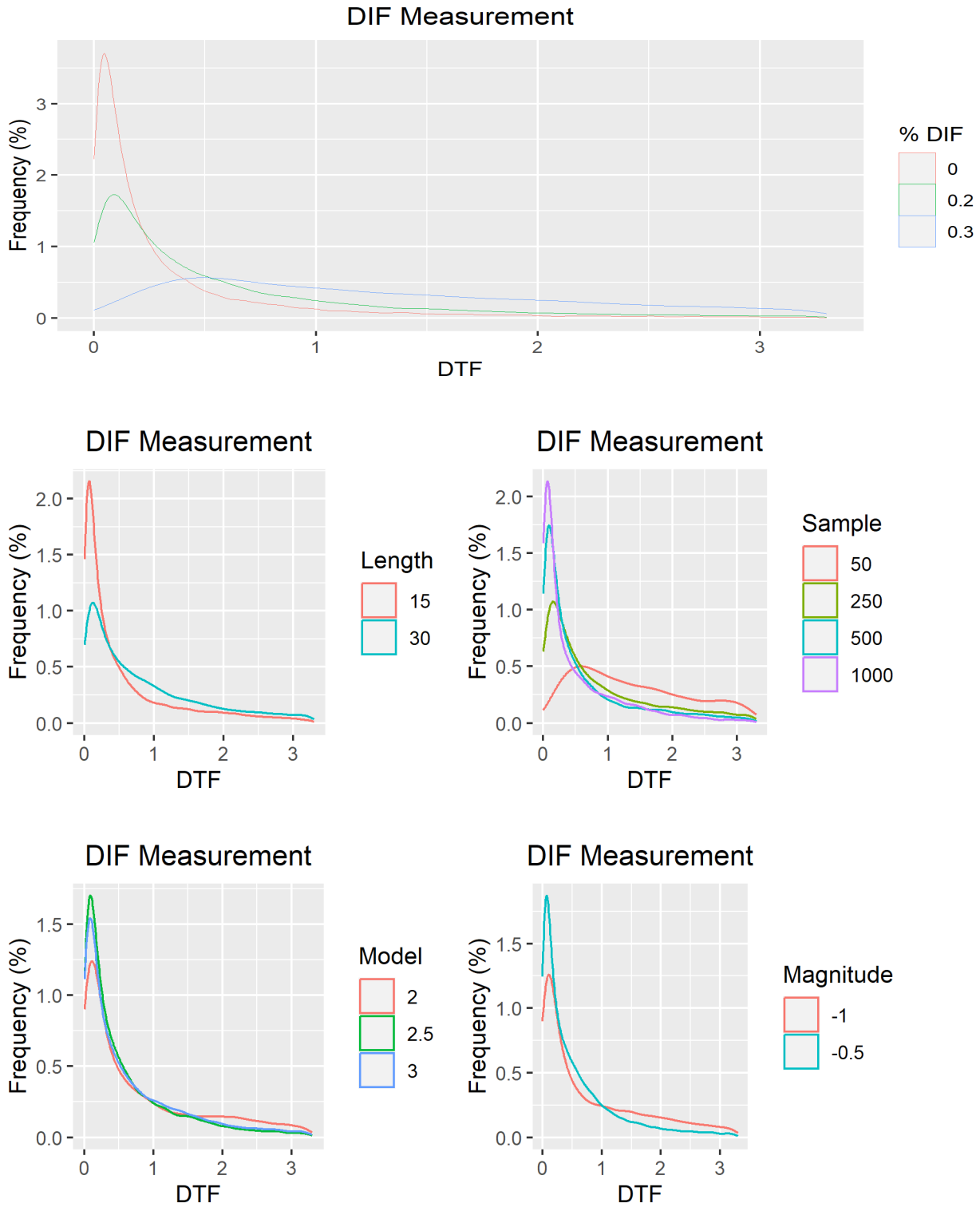
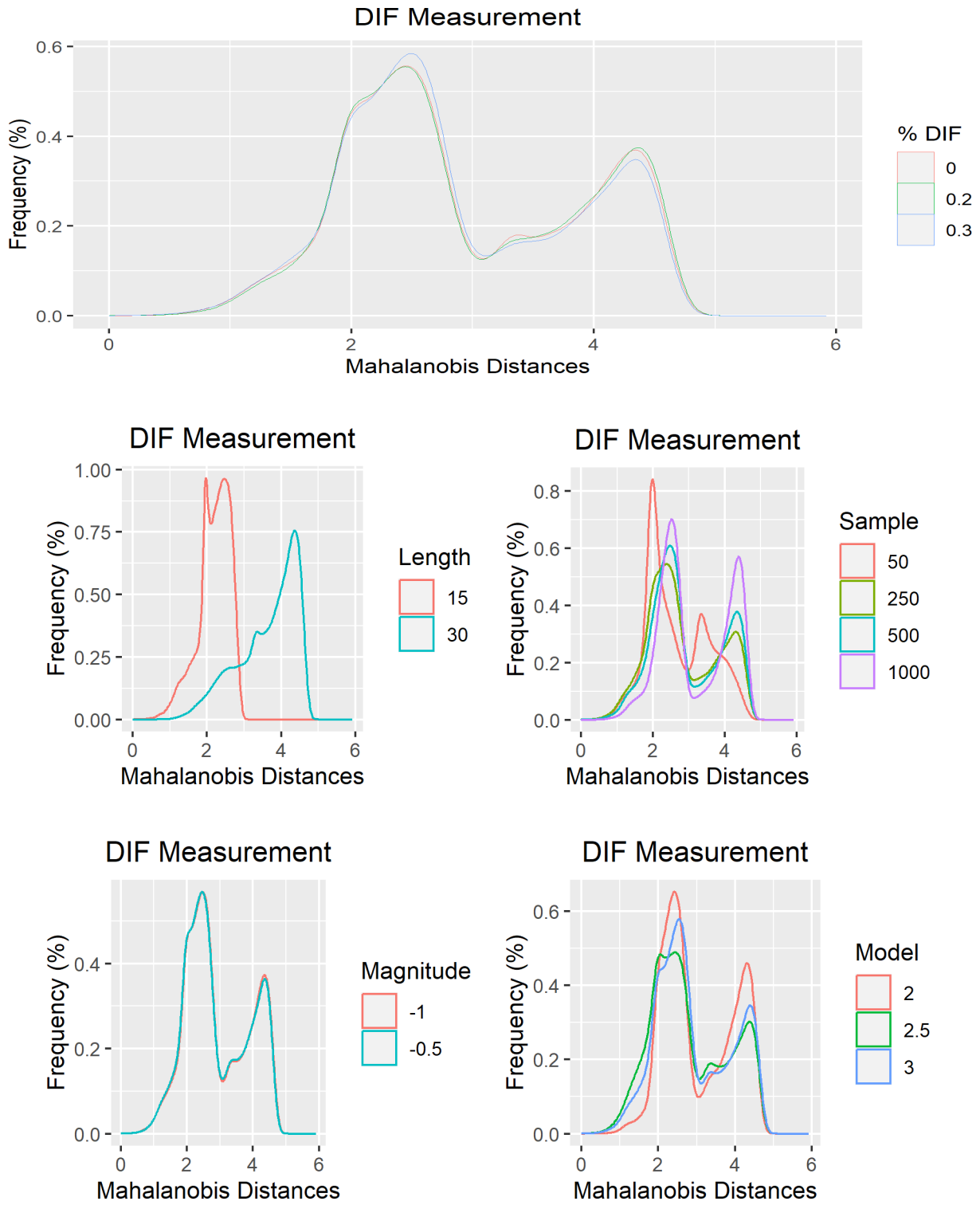


Figure 19

Mahalanobis distance performance within various conditions



Appendix C: Code Walkthrough

The study was conducted entirely within R, run via SLURM, on a supercomputer environment. The SLURM portion was straightforward and based on documentation from the local array, and the R code was factored to take advantage of this environment. Future researchers may benefit from changes to the code designed to minimize the proportion of data kept in memory during the course of the simulation; the code could be reasonably modified to do that by using segments rather than all conditions at once. The heavy lifting of the simulation is performed by the *parallel* R library; lists are passed to the *parLapply* function to execute code on threads with a minimum amount of memory overhead. Following in this appendix are block quotes of code with explanations of why it was used, how it performs, and where future researchers might want to further develop the approach used.

The code begins with some initialization and errors checking. It was designed to execute on local and supercomputer environments by tailoring initial variables not unlike environmental variables, and future researchers would be well-rewarded by refactoring this code to inherit environmental variables from SLURM or whatever workload manager is in use:

```
VERSION <- "A"
SEGMENT <- 1
NUM_NODES <- 7

conditionGenerate <- function(
  testLength = stop("testLength must be defined."),
  sampleSize = stop("sampleSize must be defined."),
  difPercent = stop("difPercent must be defined."),
  impactAmount = stop("impactAmount must be defined."),
  modelType = stop("modelType must be defined."),
  linkingAccuracy = stop("linkingAccuracy must be defined."),
  groupingAccuracy = stop("groupingAccuracy must be defined."),
  replications = stop("replications must be defined.")
) {
  if (!testLength %in% c(15, 30)) stop(paste("Bad testLength parameter:",testLength))
  if (!impactAmount %in% c(0, -.5, -1)) stop(paste("Bad impactAmount parameter:",impactAmount))
  if (!difPercent %in% c(0, .2, .3)) stop(paste("Bad difPercent parameter:",difPercent))
  if (!modelType %in% c(2, 2.5, 3)) stop(paste("Bad modelType parameter:",modelType))

  if (!sampleSize %in% c(50, 250, 500, 1000)) stop(paste("Bad sampleSize parameter:",sampleSize))
  if (!linkingAccuracy %in% c(0, 1)) stop(paste("Bad linkingAccuracy parameter:",linkingAccuracy))
  if (!groupingAccuracy %in% c(0, 1)) stop(paste("Bad groupingAccuracy parameter:",groupingAccuracy))
  if (!replications %in% c(5, 80, 100, 300, 325, 625)) stop(paste("Bad replications
parameter:",replications))
```

```
  irtParameters <- matrix(
  c(-0.07,0.21,0.54,-0.03,0.01,1.96,0.04,-0.09,-1.16,0.02,0.2,-0.43,-0.06,-0.34,0.05,-
0.25,0.06,0.31,0.04,0.13,0.52,-0.96,-0.79,0.37,-0.71,-0.19,0.74,-0.44,-0.17,0.53,

0.49,0.92,1.26,0.61,1.74,0.5,0.96,0.59,0.82,1.26,0.82,0.75,1.49,0.97,1.49,0.89,1.45,0.75,1.43,0.6,0.83,
0.56,0.67,0.7,1.03,0.89,1.23,0.9,1.23,0.69,

0.19,0.15,0.05,0.18,0.12,0.12,0.13,0.18,0.17,0.11,0.07,0.15,0.09,0.12,0.12,0.15,0.07,0.18,0.08,0.22,0.0
9,0.19,0.2,0.18,0.14,0.21,0.06,0.18,0.12,0.17),
  nrow = 30,
  dimnames = list(paste("item", 1:30),c("b","a","c"))
  )
  #Birnbaum 1969 Paramaterization:
  #c + ((1-c) /(1+exp(-1.7*a*(theta-b))))
  if (modelType == 2) {
    irtModel <- function(b, a, c, theta) return(0 + ((1-0) /(1+exp(-1.7*a*(theta-b)))))
  } else if (modelType == 2.5) {
    irtModel <- function(b, a, c, theta) return(0.2 + ((1-0.2) /(1+exp(-1.7*a*(theta-b)))))
  } else if (modelType == 3) {
    irtModel <- function(b, a, c, theta) return(c + ((1-c) /(1+exp(-1.7*a*(theta-b)))))
  } else {
    stop(paste("Bad modelType",modelType))
  }


  if (testLength == 15) {
    referenceParameters <- irtParameters[1:15,]
    focalParameters <- irtParameters[1:15,]

  } else if (testLength == 30) {
    referenceParameters <- irtParameters[1:30,]
    focalParameters <- irtParameters[1:30,]
  } else {
    stop(paste("Bad testLength:",testLength))
  }

  if (difPercent == .3) {
    if (testLength == 30) {
      focalParameters[c("item 2", "item 10", "item 17", "item 25"), "b"] <- focalParameters[c("item 2",
"item 10", "item 17", "item 25"), "b"] + impactAmount
      focalParameters[c("item 4", "item 7", "item 13", "item 19", "item 22", "item 28"), "b"] <-
focalParameters[c("item 4", "item 7", "item 13", "item 19", "item 22", "item 28"), "b"] + impactAmount
    } else {
      focalParameters[c("item 2", "item 10"), "b"] <- focalParameters[c("item 2", "item 10"), "b"] +
impactAmount
      focalParameters[c("item 4", "item 7", "item 13"), "b"] <- focalParameters[c("item 4", "item 7",
"item 13"), "b"] + impactAmount
    }
  } else if (difPercent == .2) {
    if (testLength == 30) {
      focalParameters[c("item 2", "item 10", "item 17", "item 25"), "b"] <- focalParameters[c("item 2",
"item 10", "item 17", "item 25"), "b"] + impactAmount
    } else {
      focalParameters[c("item 2", "item 10"), "b"] <- focalParameters[c("item 2", "item 10"), "b"] +
impactAmount
    }
  } else if (difPercent == 0) {
    focalParameters <- referenceParameters
  } else {
    stop(paste("Bad difPercent:",difPercent))
  }

  persons <- matrix(c(rnorm(sampleSize), rep(0, times = sampleSize)), nrow = sampleSize, dimnames =
list(paste("person",1:sampleSize), c("theta","focalGroup")))
  persons[sample(1:sampleSize, sampleSize * 0.25), "focalGroup"] <- 1

  instruments <- lapply(as.list(1:replications), function(i) {t(apply(persons, 1, function(me) {
    if (me["focalGroup"] == 1) {
```

```
      myTest <- focalParameters
    } else {
      myTest <- referenceParameters
    }
    myPs <- apply(myTest, 1, function(myItem) {
      irtModel(myItem["b"], myItem["a"], myItem["c"], me["theta"])
    })
    as.numeric(sapply(myPs, function(myP) runif(1) < myP))
  })})

  results <- list(
    instruments = instruments,
    persons = persons,
    referenceParameters = referenceParameters,
    focalParameters = focalParameters,
    condition = list(
      difPercent = difPercent,
      groupingAccuracy = groupingAccuracy,
      impactAmount = impactAmount,
      linkingAccuracy = linkingAccuracy,
      modelType = modelType,
      replications = replications,
      sampleSize = sampleSize,
      testLength = testLength
    )
  )

  return(results)
}
```

The above function is designed to prepare a list object representing a condition containing multiple repetitions of simulated data based on parameters passed to the function; this approach was used to take advantage of the nature of the *parLapply* function and to minimize the proliferation of variables in the workspace. Thus, the simulation itself can be prepared:

```
conditions <- expand.grid(testLength = c(15, 30),
                          impactAmount = c(-.5, -1),
                          difPercent = c(0, .2, .3),
                          modelType = c(2, 2.5, 3),
                          sampleSize = c(50, 250, 500, 1000),
                          linkingAccuracy = 1,
                          groupingAccuracy = 1,
                          replications = 625
                          )
```

The *expand.grid* function will prepare all permutations of these variables which then are easy to pass as arguments to *parLapply*. The following code has one significant flaw, however: the index, rather than the list itself, is the argument to the function. This necessitated the export of the *condition* and *conditionGenerate* objects to each cluster. Future researchers would benefit from refactoring this code to directly pass the *condition* object as the argument to *parLapply* and avoid linearly-scaling memory costs for running this part in parallel:

```
library(parallel)
cl <- makeCluster(NUM_NODES)
clusterExport(cl, c("conditions", "conditionGenerate"))
simulatedData <- parLapply(cl, 1:nrow(conditions), function(i) conditionGenerate(
                           testLength = conditions[i, "testLength"],
                           impactAmount = conditions[i, "impactAmount"],
                           difPercent = conditions[i, "difPercent"],
                           modelType = conditions[i, "modelType"],
                           sampleSize = conditions[i, "sampleSize"],
                           linkingAccuracy = conditions[i, "linkingAccuracy"],
                           groupingAccuracy = conditions[i, "groupingAccuracy"],
                           replications = conditions[i, "replications"]))

stopCluster(cl)
rm(cl)
save("simulatedData", file = paste0("simulatedData-",VERSION,"-",SEGMENT,"-",Sys.Date(),".RData"))
print(paste("Data simulated",Sys.time()))
gc()
```

In the study, these simulated data were saved as part of the development process and to allow for debugging in environments where there may not be enough memory to load the entire workspace the simulation will create. More confident researchers might be able to avoid this step, but it is not recommended unless disk space is at a restrictive premium or unavailable at practical speeds. Previous studies, such as Fikis and Oshima (2017), have found that slow disk space operations present significant bottlenecks in IRTPRO, but the issue should not be present with low file counts such as in this study's approach.

After this, all that remains is to prepare the fitted models and extract some diagnostic statistics from them. That, however, is a monumental task. Although the list-based structure of objects in the workspace will play well into *parLapply*, the resulting scenario is a list per condition with a list of repetitions embedded inside: nested loops will be used. Future studies with widely varied conditions might experience load balancing issues, but they were not noticeable in this study. Additionally, this study's error checking is somewhat excessive and some calculations are redundant. There are many opportunities for wresting more efficiency from R. Again, the technique was to iterate through lists and prepare objects. The *tryCatch* function does a great deal of work here preventing errors from aborting the simulation:

```
cl <- makeCluster(NUM_NODES)
analyzedData <- parLapply(cl, simulatedData, function(conditionData) {
  library(ltm)
```

```
  library(e1071)
  library(DFIT)
  models <- lapply(conditionData[["instruments"]], function(instrument) {
    focalMembers <- as.logical(conditionData[["persons"]][, "focalGroup"])

    if (conditionData[["condition"]][["groupingAccuracy"]] == 1) {
      reference <- instrument[!focalMembers, ]
      focal <- instrument[focalMembers, ]
    } else {
      kludge <- sample(1:length(focalMembers), length(focalMembers)*.05)
      focalMembers[kludge] <- !focalMembers[kludge]

      reference <- instrument[!focalMembers, ]
      focal <- instrument[focalMembers, ]
    }
    timer <- Sys.time()
    fullModel <- tryCatch(ltm(instrument ~ z1, IRT.param = TRUE), error = function(i) return(NA))
    fullModelTime <- Sys.time() - timer
    timer <- Sys.time()
    referenceModel <- tryCatch( ltm(reference ~ z1, IRT.param = TRUE) , error = function(i) return(NA))
    referenceModelTime <- Sys.time() - timer

    if (!any(is.na(referenceModel))) {
      goodAnchors <- c(1,3,5,6,8,9,11,12,14,15,16,18,20,21,23,24,26,27,29,30)
      badAnchors <- c(2,10,17,25)

      goodAnchors <- goodAnchors[goodAnchors <= conditionData[["condition"]]["testLength"]]
      badAnchors <- badAnchors[badAnchors  <= conditionData[["condition"]]["testLength"]]

      if (conditionData[["condition"]]["linkingAccuracy"] == 1) {
        myAnchors <- sample(goodAnchors, 3)
      } else {
        myAnchors <- c(sample(goodAnchors, 2), sample(badAnchors, 1))
      }

      myConstraints <- rbind(cbind(item = myAnchors, param = 1, value =
referenceModel[["coefficients"]][myAnchors, 1]),
                             cbind(item = myAnchors, param = 2, value =
referenceModel[["coefficients"]][myAnchors, 2])
      )
      timer = Sys.time()
      focalModel <- tryCatch( ltm(focal ~ z1, IRT.param = TRUE, constr = myConstraints) , error =
function(i) return(NA))
      focalModelTime <- Sys.time() - timer

    } else {
      focalModel <- NA
      focalModelTime <- NA
    }

    if (!any(is.na(focalModel))) {
      dtfParams = tryCatch( list(
                              reference = cbind(

summary(referenceModel)[["coefficients"]][(conditionData[["condition"]][["testLength"]])+(1:conditionDa
ta[["condition"]][["testLength"]])],

summary(referenceModel)[["coefficients"]][1:conditionData[["condition"]][["testLength"]]]
                              ),
                              focal = cbind(

summary(focalModel)[["coefficients"]][(conditionData[["condition"]][["testLength"]])+(1:conditionData[[
"condition"]][["testLength"]])],

summary(focalModel)[["coefficients"]][1:conditionData[["condition"]][["testLength"]]]
                              )
                            ),
                            error = function(i) return(-666))
      if (!any(is.na(dtfParams))) {
```

```
        timer <- Sys.time()
        myDTF <- tryCatch(Dtf(itemParameters = dtfParams, irtModel = "2pl"), error = function(i)
return(-777))
        myDTFTime <- Sys.time() - timer
      } else {
        myDTF <- -888
        myDTFTime <- NA
      }
    } else {
      myDTF <- -999
      myDTFTime <- NA
    }

    if (!any(is.na(fullModel))) {
      fullErrs <- tryCatch( matrix(summary(fullModel)[["coefficients"]][, "std.err"], nrow =
conditionData[["condition"]][["testLength"]]) , error = function(i) return(NA))

      if (!any(is.na(fullErrs))) {
        timer <- Sys.time()
        mahalas <- tryCatch( mahalanobis(fullErrs, center = colMeans(fullErrs), cov = cov(fullErrs)) ,
error = function(i) return(NA))
        mahalasTime <- Sys.time() - timer
      } else {
        mahalas <- NA
        mahalasTime <- NA
      }

      if (!any(is.na(mahalas))) {
        newStat <- tryCatch( skewness(mahalas), error = function(i) return(NA))
        newStatMean <- tryCatch( mean(mahalas), error = function(i) return(NA))
        newStatMedian <- tryCatch( median(mahalas), error = function(i) return(NA))
        newStatSd <- tryCatch( sd(mahalas), error = function(i) return(NA))
        newStatChiSq <- tryCatch( sum(mahalas), error = function(i) return(NA))
        newStatChiSqCrit <- tryCatch( qchisq(p = .05, df = (length(mahalas) - 1), lower.tail= FALSE),
error = function(i) return(NA))
        newStatChiSqTest <- tryCatch( as.numeric(newStatChiSq >= newStatChiSqCrit), error = function(i)
return(NA))
      } else {
        newStat <- NA
        newStatMean <- NA
        newStatMedian <- NA
        newStatSd <- NA
        newStatChiSq <- NA
        newStatChiSqCrit <- NA
        newStatChiSqTest <- NA
      }
    } else {
      fullErrs <- NA
      mahalas <- NA
      newStat <- NA
      newStatMean <- NA
      newStatMedian <- NA
      newStatSd <- NA
      newStatChiSq <- NA
      newStatChiSqCrit <- NA
      newStatChiSqTest <- NA
    }

    analysis <- list(
      fullModel = fullModel,
      referenceModel = referenceModel,
      myAnchors = myAnchors,
      focalModel = focalModel,
      myDTF = myDTF,
      fullErrs = fullErrs,
      mahalas = mahalas,
      newStat = newStat,
      newStatMean = newStatMean,
      newStatMedian = newStatMedian,
```

```
      newStatSd = newStatSd,
      newStatChiSq = newStatChiSq,
      newStatChiSqCrit = newStatChiSqCrit,
      newStatChiSqTest = newStatChiSqTest,
      fullModelTime = fullModelTime,
      referenceModelTime = referenceModelTime,
      focalModelTime = focalModelTime,
      myDTFTime = myDTFTime,
      mahalasTime = mahalasTime
    )

    return(analysis)
  })

  conditions <- matrix(rep(unlist(conditionData[["condition"]]), times = length(models)), nrow =
length(models), byrow = TRUE)
  conditions <- cbind(conditions, 1:length(conditions))


  newStats <- unlist(lapply(models, function(i) i[["newStat"]]))
  newStatMeans <- unlist(lapply(models, function(i) i[["newStatMean"]]))
  newStatMedians <- unlist(lapply(models, function(i) i[["newStatMedian"]]))
  newStatSds <- unlist(lapply(models, function(i) i[["newStatSd"]]))
  DTFs <- unlist(lapply(models, function(i) i[["myDTF"]]))
  newStatChiSqs <- unlist(lapply(models, function(i) i[["newStatChiSq"]]))
  newStatChiSqCrits <- unlist(lapply(models, function(i) i[["newStatChiSqCrit"]]))
  newStatChiSqTests <- unlist(lapply(models, function(i) i[["newStatChiSqTest"]]))

  fullModelTimes <- unlist(lapply(models, function(i) i[["fullModelTime"]]))
  referenceModelTimes <- unlist(lapply(models, function(i) i[["referenceModelTime"]]))
  focalModelTimes <- unlist(lapply(models, function(i) i[["focalModelTime"]]))
  myDTFTimes <- unlist(lapply(models, function(i) i[["myDTFTime"]]))
  mahalasTimes <- unlist(lapply(models, function(i) i[["mahalasTime"]]))

  conditions <- cbind(conditions, newStats, DTFs, newStatMeans, newStatMedians, newStatSds,
newStatChiSqs, newStatChiSqCrits, newStatChiSqTests,
                      fullModelTimes, referenceModelTimes, focalModelTimes, myDTFTimes, mahalasTimes)
  colnames(conditions) <- c(names(unlist(conditionData[["condition"]])),"rep","newStat","DTF",
"newStatMeans", "newStatMedians", "newStatSds", "newStatChiSq", "newStatChiSqCrit", "newStatChiSqTest",
                            "timeFull", "timeReference", "timeFocal", "timeDTF", "timeMahalanobis" )

  results <- list(
    simulation = conditions,
    models = models
  )
  return(results)
})

stopCluster(cl)
rm(cl)

save("analyzedData", file = paste0("analyzedData-",VERSION,"-",SEGMENT,"-",Sys.Date(),".RData"))
print(paste("Data analyzed",Sys.time()))
gc()
cl <- makeCluster(NUM_NODES)
# clusterExport(cl, c("analyzedData"))
simulationOutput <- parLapply(cl, analyzedData, function(i) return(i[["simulation"]]))
stopCluster(cl)
rm(cl)
simulationOutput <- do.call(rbind, simulationOutput)
save("simulationOutput", file = paste0("simulationOutput-",VERSION,"-",SEGMENT,"-
",Sys.Date(),".RData"))
print(paste("Model Extracted",Sys.time()))
gc()
```

Some of the functionality of this code, such as linking and grouping accuracy conditions, wound up unused in the study. Importantly, R packages such as *ltm* and *DFIT* needed to be

loaded within the *parLapply* function so that each thread loaded the library; this was a common mistake in very early iterations of the code. The *psych* package's function for measuring skewness was found to give equivalent values to that in the *e1071* package. Some statistics were not found suitable for analysis, and, technically, the original simulated data in the results list is redundant in convergent conditions; an *ltm* object contains the instrument it was fit with in its *x* object. The results maintained this object, though, for ease of extraction and retention in the case of failed models. The output data loops to ensure that the *analyzedData* object is a matrix with minimized filesize suitable for exploration on personal computers. Future developments for this code may include bringing in evaluation of NCDIF, which was prepared in a subsequent post-hoc analysis:

```
findings <- parLapply(cl, analyzedData, function(thisCondition) {
  library(ltm)
  paramExtract <- function(WHICH_MODEL = stop("Specify which model: fullModel, referenceModel,
focalModel"), thisRep = stop("Need thisRep")) {
    if (length(thisRep[[WHICH_MODEL]]) == 1) {
      paramModelB <- rep(NA, times = 30)
      paramModelA <- rep(NA, times = 30)
      modelMahalas <- NA
    } else {
      modelParams <- tryCatch(
        summary(thisRep[[WHICH_MODEL]])[["coefficients"]][ , "value"],
        error = function(i) return(NA)
      )
      modelErrors <- tryCatch(
        summary(thisRep[[WHICH_MODEL]])[["coefficients"]][ , "std.err"],
        error = function(i) return(NA)
      )
      if (!any(is.na(modelErrors))) {
        if (WHICH_MODEL == "fullModel") {
          aErrors <- modelErrors[grepl("Dscrmn", names(modelErrors))]
          bErrors <- modelErrors[grepl("Dffclt", names(modelErrors))]
          modelErrors <- matrix(modelErrors, nrow = length(modelErrors)/2)
          uErrors <- tryCatch(apply(modelErrors, 1, function(i) {dist(rbind(i, c(0,0)))}), error =
function(i) return(NA))
          uzErrors <- tryCatch(apply(modelErrors, 1, function(i) {dist(rbind(
            c( ((i[1]-mean(modelErrors[,1], na.rm=TRUE)) /sd(modelErrors[,1], na.rm=TRUE)),
               ((i[2]-mean(modelErrors[,2], na.rm=TRUE)) /sd(modelErrors[,2], na.rm=TRUE))
            ), c(0,0)))}), error = function(i) return(NA))
        } else {
          modelErrors <- matrix(modelErrors, nrow = length(modelErrors)/2)
          aErrors <- NA
          bErrors <- NA
          uErrors <- NA
          uzErrors <- NA
        }
        modelMahalas <- tryCatch( mahalanobis(modelErrors, center = colMeans(modelErrors), cov =
cov(modelErrors)) , error = function(i) return(NA))
      } else {
        modelMahalas <- NA
```

```
        aErrors <- NA
        bErrors <- NA
        uErrors <- NA
        uzErrors <- NA
      }
      aErrors <- c(aErrors, rep(NA, times = (30-length(aErrors))))
      bErrors <- c(bErrors, rep(NA, times = (30-length(bErrors))))
      uErrors <- c(uErrors, rep(NA, times = (30-length(uErrors))))
      uzErrors <- c(uzErrors, rep(NA, times = (30-length(uzErrors))))
      names(aErrors) <- paste0(WHICH_MODEL, "ErrorA", formatC(1:30, width = 2, format = "d", flag =
"0"))
      names(bErrors) <- paste0(WHICH_MODEL, "ErrorB", formatC(1:30, width = 2, format = "d", flag =
"0"))
      names(uErrors) <- paste0(WHICH_MODEL, "ErrorU", formatC(1:30, width = 2, format = "d", flag =
"0"))
      names(uzErrors) <- paste0(WHICH_MODEL, "ErrorUZ", formatC(1:30, width = 2, format = "d", flag =
"0"))
      if (length(modelParams) == 30) {
        paramModelB <- c(modelParams[1:15], rep(NA, times = 15))
        paramModelA <- c(modelParams[16:30], rep(NA, times = 15))
      } else {
        if (length(modelParams) == 60) {
          paramModelB <- modelParams[1:30]
          paramModelA <- modelParams[31:60]
        } else {
          paramModelB <- rep(NA, times = 30)
          paramModelA <- rep(NA, times = 30)
        }
      }
    }
    if (all(is.na(modelMahalas))) {
      maxMahalas <- NA
      minMahalas <- NA
    } else {
      maxMahalas <- max(modelMahalas, na.rm = TRUE)
      minMahalas <- min(modelMahalas, na.rm = TRUE)
    }
    #modelMahalas <<- modelMahalas

    if (length(modelMahalas) != 30) {
      modelMahalas <- c(modelMahalas, rep(NA, times = (30 - length(modelMahalas))))
    }
    names(modelMahalas) <- paste0(WHICH_MODEL, "M", formatC(1:30, width = 2, format = "d", flag = "0"))
    names(paramModelB) <- paste0(WHICH_MODEL, "B", formatC(1:30, width = 2, format = "d", flag = "0"))
    names(paramModelA) <- paste0(WHICH_MODEL, "A", formatC(1:30, width = 2, format = "d", flag = "0"))

    paramModelB <- paramModelB
    paramModelA <- paramModelA
    names(maxMahalas) <- paste0(WHICH_MODEL, "MaxMahalanobis")
    names(minMahalas) <- paste0(WHICH_MODEL, "MinMahalanobis")

    if (WHICH_MODEL == "fullModel") {
      return(c(maxMahalas, minMahalas, modelMahalas, paramModelB, paramModelA, aErrors, bErrors,
uErrors, uzErrors))
    } else {
      return(c(paramModelB, paramModelA))
    }


  }
  conditionParams <- thisCondition[["simulation"]]

  conditionOutcomes <- lapply(thisCondition[["models"]], function(thisRep) {
      focalModelParams <-
tryCatch( matrix(summary(thisRep[["focalModel"]])[["coefficients"]][,"value"], nrow =
length(summary(thisRep[["focalModel"]])[["coefficients"]][,"value"])/2)[, c(2,1)] , error = function(i)
return(NA))
      referenceModelParams <-
tryCatch( matrix(summary(thisRep[["referenceModel"]])[["coefficients"]][,"value"], nrow =
```

```
length(summary(thisRep[["referenceModel"]])[["coefficients"]][,"value"])/2)[, c(2,1)] , error =
function(i) return(NA))


    focalThetas <- tryCatch( factor.scores(thisRep[["focalModel"]])[[1]][, "z1"] , error = function(i)
return(NA))
    trialNCDIFS <- tryCatch( Ncdif(list(focal = focalModelParams, reference = referenceModelParams),
irtModel = "2pl", focalAbilities = focalThetas) , error = function(i) return(NA))

    if (length(trialNCDIFS) != 30) {
      trialNCDIFS <- c(trialNCDIFS, rep(NA, times = (30-length(trialNCDIFS))))
    }
    names(trialNCDIFS) <- paste0("ncdif", formatC(1:30, width = 2, format = "d", flag = "0"))

    c(paramExtract("fullModel", thisRep), paramExtract("referenceModel", thisRep),
paramExtract("focalModel", thisRep), trialNCDIFS)
  })
  return(cbind(conditionParams, do.call(rbind, conditionOutcomes)))
})
stopCluster(cl)
rm(cl)

findings <- do.call(rbind, findings)


save("findings", file = paste0("findings-",VERSION,"-",SEGMENT,"-",Sys.Date(),".RData"))
```

This post-hoc analysis was some of the most inefficient code, but did not need further optimization for the purposes of this study. Try *tryCatch* function again plays an important role in ensuring the code finishes execution even in the face of failed repetitions. During debugging, it was found very useful to temporary refactor the *parLapply* as a single-threaded *lapply* function with objects pushed to the parent environment with the <<- assignment operator. R has some peculiarities with how lists are iterated and available either with single or double square brackets; testing proved essential to ensure code performed as expected in light of *tryCatch*. Those replicating this study will also find it important to recode fail values from *tryCatch* as missing.

Finally, charts and graphs were prepared with straightforward implementation of functions such as *lm*, *anova*, and charts via the *ggplot2* package for ease of configuration. Some recoding was necessary to assign appropriate *NA* values, and care was required in specifying discrimination and difficulty parameters: the *DFIT* and *ltm* packages use different parameterization warranting a great deal of caution, but applying the *summary* function to an *ltm* model called fitted with the *IRT.param* argument provided well-labelled, appropriate values.