

University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Doctoral Dissertations


Dissertations and Theses

November 2016

Identifying Examinees Who Possess Distinct and Reliable Subscores When Added Value is Lacking for the Total Sample

Joseph A. Rios
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Quantitative Psychology Commons](#), and the [Statistical Methodology Commons](#)

Recommended Citation

Rios, Joseph A., "Identifying Examinees Who Possess Distinct and Reliable Subscores When Added Value is Lacking for the Total Sample" (2016). *Doctoral Dissertations*. 800.
https://scholarworks.umass.edu/dissertations_2/800

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

IDENTIFYING EXAMINEES WHO POSSESS DISTINCT AND RELIABLE
SUBSCORES WHEN ADDED VALUE IS LACKING FOR THE TOTAL SAMPLE

A Dissertation Presented

by

JOSEPH A. RIOS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Education

© Copyright by Joseph A. Rios 2016

All Rights Reserved

IDENTIFYING EXAMINEES WHO POSSESS DISTINCT AND RELIABLE
SUBSCORES WHEN ADDED VALUE IS LACKING FOR THE TOTAL SAMPLE

A Dissertation Presented

By

JOSEPH A. RIOS

Approved as to style and content by:

Stephen G. Sireci, Chair

Craig S. Wells, Member

Michael Lavine, Member

Joseph Berger, Senior Associate Dean
College of Education

DEDICATION

I would like to dedicate this dissertation to my family and friends for their consistent love, support, and inspiration.

ACKNOWLEDGMENTS

I would like to thank the faculty within the Research, Educational Measurement, and Psychometrics program at the University of Massachusetts Amherst for providing an environment that stimulated my learning on a daily basis.

ABSTRACT

IDENTIFYING EXAMINEES WHO POSSESS DISTINCT AND RELIABLE SUBSCORES WHEN ADDED VALUE IS LACKING FOR THE TOTAL SAMPLE

SEPTEMBER 2016

JOSEPH A. RIOS, B.A., LEWIS & CLARK COLLEGE
M.A., UNIVERSITY OF CALIFORNIA, RIVERSIDE
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Dr. Stephen G. Sireci

Research has demonstrated that although subdomain information may provide no added value beyond the total score, in some contexts such information is of utility to particular demographic subgroups (Sinharay & Haberman, 2014). However, it is argued that the utility of reporting subscores for an individual should not be based on one's manifest characteristics (e.g., gender or ethnicity), but rather on individual needs for diagnostic information, which is driven by multidimensionality in subdomain scores. To improve the validity of diagnostic information, this study proposed the use of Mahalanobis Distance and H^T indices to assess whether an individual's data significantly departs from unidimensionality. Those examinees that were found to differ significantly were then assessed separately for subscore added value via Haberman's (2008) procedure. To this end, simulation analyses were conducted to evaluate Type I error, power, and recovery of subscore added value classifications for various levels of subdomain test lengths, subdomain inter-correlations, and proportions of multidimensionality in the total sample. Results demonstrated that the H^T index possessed around 100% power across all conditions, while maintaining Type I error below 5%,

which led to nearly perfect recovery of subscore added value classifications. In contrast, the power rates for Mahalanobis Distance were much lower ranging from 13% to 61% with Type I errors maintained at the nominal level of 5%. Although the power rates were below the desired criterion of 80%, the cases identified as aberrant using this method were found to have greater variability between subdomain scores, increased reliability, and lower observed subdomain correlations when compared to the generated data. As a result, outlier cases were found to have subscore added value for nearly 100% of cases across conditions even when the generated multidimensional data did not possess subscore added value. These results were cross-validated using a large-scale high-stakes test in which the Mahalanobis Distance measure was found to identify 6.57% of 8,803 test-takers that possessed subscores with added-value who otherwise would have been masked by the unidimensionality of the total sample. Overall, this study suggests that the Mahalanobis Distance measure shows some promise in identifying examinees with multidimensional score profiles.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
ABSTRACT	vi
LIST OF TABLES	xii
LIST OF FIGURES	xiii
CHAPTER	
1. INTRODUCTION	1
1.1 Background	1
1.2 Impact of Diagnostic Score Reporting on Instruction and Test Performance	3
1.3 Statement of Problem.....	7
1.4 Purpose of Study	9
2. REVIEW OF LITERATURE	12
2.1 Overview of Literature Review	12
2.2 Review of Subscore Estimation Methodologies	13
2.2.1 Simple Subscore Estimation Methods.	14
2.2.1.1 Number-Correct or Percent-Correct Raw Subscores.....	14
2.2.1.2 Simple Unidimensional IRT Subscore Estimation Procedures	15
2.2.2 Subscore Augmentation Estimation Methodologies.....	15
2.2.2.1 Kelley’s (1947) Univariate Regression.....	15
2.2.2.2 Wainer et al.’s (2001) Subscore Augmentation.....	16
2.2.2.3 Wainer et al.’s (2001) Augmentation Method with IRT Theta Estimates.....	18
2.2.2.4 Objective Performance Index.	19

2.2.2.5 Out-of-Scale Information Method	21
2.2.2.6 Multidimensional IRT Applications to Subscore Estimation	22
2.2.2.6.1 Compensatory MIRT Models	23
2.2.2.6.2 Non-Compensatory MIRT Models	24
2.2.2.6.3 Cognitive Diagnostic Models.	25
2.3 Comparative Analyses of Subscore Estimation Methodologies	26
2.3.1 Comparison of Methods for Dichotomous Items.....	26
2.3.2 Comparison of Methods for Mixed Format Tests.....	30
2.4 Subscore Reporting in Practice.....	35
2.5 Validation of Subscore Reporting in Practice.....	38
2.5.1 Methods for Assessing the Adequacy of Reporting Raw Subscores.	39
2.5.1.1 Haberman’s (2008) Method.....	39
2.5.1.2 Added Value Based on Classifications.	41
2.5.1.3 Assessing Invariance of Subscore Added Value	43
2.5.1.4 Added Value of Reporting Subscores at the Aggregate-Level.	44
2.5.1.5 Brennan’s (2012) Utility Index.....	46
2.5.2 Studies that Have Evaluated Subscore Validity for Operational Data.	48
2.5.2.1 Individual-Level Subscore Validity.....	48
2.5.2.2 Subscore Added Value Invariance.....	53
2.6 Summary of Literature Review.....	56
3. METHODOLOGY	60
3.1 Overview.....	60
3.2 Simulation Study.....	61

3.2.1 Data Generation.....	61
3.2.2 Degree of Masking Effects when Assessing Subscore Added Value.....	63
3.2.3 Assessment of Aberrant Score Profiles and Patterns at the Individual-Level.....	63
3.2.3.1 Identification of Multivariate Outlier Cases.....	64
3.2.3.2 Person-Fit Indices.....	66
3.2.4 Added Value Evaluation by Outlier Classification.....	69
3.2.5 Independent Variables.....	69
3.2.6 Dependent Variables.....	72
3.2.6.1 Type I Error.....	72
3.2.6.2 Power.....	72
3.2.6.3 Recovery of Subscore Added Value Classifications.....	73
3.3 Application of Aberrant Identification Methods to Real Data.....	73
4. RESULTS.....	77
4.1 Overview of Results Section.....	77
4.2 Degree of Masking Effects when Assessing Subscore Added Value.....	77
4.2.1 Conditions for Added Value of Multidimensional Generated Data.....	77
4.2.2 Masking Effects for Minority Percentages of Multidimensional Scores.....	78
4.3 Type I Error and Power by Aberrant Response Identification Procedure.....	80
4.3.1 Type I Error.....	80
4.3.2 Power.....	81
4.3.3 Recovery of Subscore Added Value Classifications.....	82
4.4 Application of Aberrant Identification Methods to Real Data.....	87

4.4.1 Profile Analysis.....	88
4.4.2 Subscore Added Value Analysis.....	89
5. DISCUSSION.....	111
5.1 Overview.....	111
5.2 Degree of Masking Effects on Subscore Added Value	112
5.3 Aberrant Detection Procedures for Assessing Subscore Added Value Invariance.....	117
5.4 Limitations and Directions for Future Research.....	121
5.5 Conclusion	127
REFERENCES	133

LIST OF TABLES

Table	Page
1. Subscore added value for generating multidimensional data by condition	94
2. Subscore added value for total sample by subdomain test length, proportion of outliers, and inter-subdomain correlations.....	95
3. Type I error rate by aberrant detection procedure	96
4. Power rate by aberrant detection procedure.....	97
5. Confusion matrix for the H^T condition with the highest power rate.....	98
6. Confusion matrix for the H^T condition with the lowest power rate	98
7. Confusion matrix for the Mahalanobis Distance condition with the lowest power rate.....	99
8. Confusion matrix for the Mahalanobis Distance condition with the highest power rate.....	99
9. Bias of PRMSE values and percentage of subscore added value for outlier groups identified by individual indices.....	100
10. Bias of descriptive statistics for aberrant responders identified using Mahalanobis Distance.....	101
11. Analysis of added value by identification of aberrant responses for applied data....	102
12. Profile analysis of applied data for total sample and outliers by index	103
13. Mean subdomain differences between aberrant and non-aberrant cases by detection method.....	104

LIST OF FIGURES

Figure	Page
1. Multidimensional cases incorrectly identified as non-aberrant from the unidimensional model by the Mahalanobis Distance measure.	105
2. Multidimensional cases correctly identified as non-aberrant from the unidimensional model by the Mahalanobis Distance measure.	106
3. Comparison of mean score profile variability and standard error of measurement for outlier observations identified via the Mahalanobis Distance measure across conditions.....	107
4. Scatterplot of outlier classifications by the Mahalanobis Distance and H^T indices for the applied dataset.	108
5. Score profiles of applied dataset for total sample and outliers identified by procedure.....	109
6. Applied comparison of mean score profile variability and standard error of measurement for outlier observations identified via the Mahalanobis Distance measure.	110
7. H^T distributions by generating dimensionality and condition.....	130
8. Distributions of the Mahalanobis Distance index by generating dimensionality and condition.....	131
9. Item characteristic curves for each item by subdomain with a subdomain test length of 10 items.....	132

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, educational assessments have increasingly assumed a central role in educational reform by serving as a measure of school accountability and teacher evaluation (Heubert & Hauser, 1998). However, it is argued that at their most basic level, educational assessments serve as a mechanism to identify student learning needs and instructional improvements. To accomplish these latter endeavors, increased research has focused on making data accessible and easy to understand, providing evidence to support feedback credibility, and leveraging technology (e.g., automated scoring) to improve the timeliness of assessment results (Coe, 1998; Hambleton & Zenisky, 2013; Marsh, Pane, & Hamilton, 2006; Smither, London, & Reilly, 2005; Yattali & Powers, 2010). In spite of this, one question looms large, what kind of feedback should be provided?

In general, there are three types of assessment feedback, which are summative, normative, and diagnostic in nature (Hambleton & Zenisky, 2013). Summative and normative information respectively communicate to the examinee and/or stakeholders how the examinee performed and how that performance is related relative to other examinees. In contrast, diagnostic information informs the examinee and/or stakeholders with information that is more detailed than that reported at the general subject area level (i.e., summative and normative information) for the purpose of informing preparation for future test administrations (Goodman & Hambleton, 2004). More specifically, diagnostic score reporting most often disseminates information related to sub-domains, which refer to a meaningful cluster of items that are based on content categories. For example, a math

test may include sub-domains on algebra, geometry, measurement, etc. The scores assigned to these subsections are generally referred to as subscores, which have become of increasing interest as the U.S. Government's No Child Left Behind Act of 2001 has demanded that students receive diagnostic reports (Sinharay, Puhan, & Haberman, 2011).

One major reason for the increased interest in subscore feedback is that it has been perceived as a component for effective teaching as it allows teachers to understand student learning challenges at a fine-grained level (Firestone, 2014; Kunnan & Jang, 2009). Elawar and Corno (1985) supported this claim by demonstrating improved student performance when providing feedback on homework that considered the following questions: "What is the key error? What is the probable reason the student made this error? How can I guide the student to avoid the error in the future? and What did the student do well that could be noted?" (p. 166). By placing focus on these types of questions when providing feedback, one can better diagnose a problem and assist in improving instruction. As an example, research has demonstrated more effective instructional targeting for low-achieving students when providing diagnostic feedback from curriculum-based measurement (CBM; Capizzi & Fuchs, 2005). One explanation for this finding is that feedback can enhance instructional competence by helping teachers recognize their accomplishments and deficiencies (Firestone, 2014). Although research has shown that diagnostic feedback has been successful on classroom-level assessments (i.e., homework and CBM) of student progress (Capizzi & Fuchs, 2005; Elewar & Corno, 1985), the next section reviews research that has investigated how diagnostic score reporting has impacted both instruction and test performance from large-scale educational assessments.

1.2 Impact of Diagnostic Score Reporting on Instruction and Test Performance

Research on feedback interventions date back over 100 years; however, surprisingly a dearth of literature exists for evaluating the impact of diagnostic feedback on instruction and test performance from large-scale assessments. More specifically, there is a lack of quantitative research that has investigated if teachers or instructors use test score feedback, how such information is implemented to improve instruction, and whether such feedback improves student performance. This is troubling as the revenue in data management and data analysis services for K-12 testing has grown exponentially to increase by \$46.2 million over a three-year period (Stein, 2003). The paragraphs that follow will describe the existing literature that has examined teacher usage of diagnostic information and its impact on improving student performance.

One of the few studies that has evaluated the use of test score feedback by teachers was conducted by Tyler (2013). In this study, teacher usage was investigated for a web-based tool implemented by the Cincinnati Public Schools to assist in the presentation and analysis of diagnostic student test scores. Within this web-based tool, teachers were able to access benchmark assessments, end-of-year state-level assessments (historical trends were available at the student-level), and for a proportion of struggling schools, pre-test and post-test data were given in September and January, respectively. These data were made available for whole classrooms, groups of students within a classroom, individual students, and at the item-level of the assessment. Furthermore, teachers were also provided resource information (e.g., lesson plans) to address the needs of struggling students based on diagnostic test score feedback. Usage of the web-based

tool was analyzed for 429 teachers in grades 3-8 who taught math, language arts, social studies, science, or a self-contained elementary classroom.

Results demonstrated that the median number of logins for teachers was 28 with median time spent viewing data online being equal to 3.5 hours through the 2008-2009 school year. However, teachers were also able to visit the web-based tool for the purpose of printing student test data. The author found that teachers on average printed group-level and individual-level test results once every three weeks and once every six weeks, respectively. On any given week only 10 to 40 percent of teachers utilized the web-based tool. For those teachers who did login, 20 to 50 percent of their time was dedicated to looking at student test performance data¹. Teachers in grades 3-6 were found to spend 38 percent less time viewing the web-based tool than their counterparts in grades 7-8. Usage was also evaluated in terms of when information was accessed. Results showed that on average teachers spent 50 percent more time viewing test results per week following a two week period after a benchmark test was administered than any other time. Interestingly, the author found that teachers spent less time during and after the state exams when compared to the benchmark test or any other time during the academic school year.

To better contextualize the results of the analysis, Tyler (2013) conducted qualitative interviews of 6 to 8 teachers from four different Cincinnati elementary schools. From these discussions, the one major contributing factor related to the lack of diagnostic feedback usage was time. For one, teachers felt that they lacked instructional time to address the student feedback provided by the assessments. Secondly, teachers

¹ The web-based tool also provided teachers with disciplinary, attendance, and grade records for individual students.

generally felt they did not have enough time to access and process the data provided. The last concern may point toward the need for school administrators to communicate how teachers should allocate time to viewing and incorporating test score feedback into instructional practices. As noted by Tyler, if the school district expects teachers to access the information in their off-time, regardless of available data analysis tools or district support, usage would not be expected to be high. Overall, results of this study are troubling as there is evidence that when data systems that provide diagnostic information are made available on a voluntary basis, teachers make very little use of them.

In contrast to Tyler (2013), Muralidharan and Sundararaman (2010) conducted an experimental study to evaluate the impact of diagnostic tests and feedback on teacher classroom behavior and student test performance in 300 randomly selected primary schools in India. Within this study, three experimental conditions each comprised of 100 schools were included: (a) no feedback, (b) feedback, and (c) feedback with monetary incentives. Across all three conditions, a diagnostic test of mathematics and language were administered at the beginning of the academic year. For the feedback schools a detailed written diagnostic score report on student performance (both absolute and relative) were provided for teachers with a personal visit from educational experts on how to read and use the performance reports and benchmarks. These schools were also made aware that end-of-year student progress would be monitored with a follow-up diagnostic test. In addition to student performance, classroom observers visited feedback schools once a month for 20-30 minutes to observe and evaluate teaching processes.

Results of the classroom observations demonstrated that when compared to the no feedback condition, teachers in feedback schools more often: taught actively, addressed

questions to students, encouraged participation, read from textbooks, made children read from textbooks, actively used the blackboard, assigned homework, provided homework guidance, provided feedback on homework, made students use a textbook, and students more often asked questions in class. Interestingly, the intervention groups did not differ in terms of teacher absence, orderly classrooms, administering tests, calling students by name, providing individual and group help, and controlling the classroom. When comparing feedback schools (incentives versus no incentives), no statistically significant differences across variables related to teaching processes were observed.

In terms of student performance, the feedback alone group did not have significantly higher scores for mathematics, language, or combined domains when compared to the no feedback group. This finding suggests that teachers within the feedback alone condition were able to model desired behaviors when observers were present in the classroom, but were not able to improve student performance beyond that achieved by the control group. Interestingly, when investigating student performance differences between feedback conditions (incentives versus no incentives), the incentive group was found to have significantly higher mean scores. This suggests that when performance-linked incentives were provided, teachers were able to more effectively utilize the diagnostic feedback for instructional purposes, due possibly to increased motivation to use such feedback. This assumption was supported by teachers in the incentives group reporting more often that feedback was useful, which was significantly correlated with student performance. Such a result implies that it is not enough to merely provide diagnostic feedback, but instead one must create an environment in which there is a demand by teachers for data-based decision making.

Although quantitative research related to data usage and more specifically diagnostic feedback are currently lacking for large-scale educational assessments, the few published studies in this area highlight a number of important points. For one, teachers are not necessarily intrinsically motivated to use diagnostic information, particularly when time allocation for interpreting and implementing such information is perceived to be limited. Secondly, diagnostic information may improve student performance when teachers are provided with training on how to use it and are given incentives for doing so. As a result, it appears that *under the right conditions* providing diagnostic information may be a worthwhile endeavor for measurement specialists; however, to ensure that instructional decisions related to detailed performance-level information are accurate, a number of psychometric concerns must be first addressed.

1.3 Statement of Problem

A matter of concern related to subscore reporting is the precision of inferences that can be made about strand-level performance, particularly as subdomains are often comprised of a small number of items or are retrofitted from previously developed unidimensional assessments. Although many stakeholders demand that subscores are reported (Brennan, 2012), the professional measurement community has warned against reporting subscores that are not of adequate psychometric quality for two reasons: 1) lack of validity evidence based on internal structure (construct validity) and 2) inadequate reliability (i.e., an inconsistency in test scores across parallel forms due largely to random measurement error). The former concern is directly addressed in Standard 1.13 of the *Standards for Educational and Psychological Testing* [American Educational Research Association (AERA), American Psychological Association (APA), & National Council

on Measurement in Education (NCME), 2014], which states that “When a test provides more than one score, the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed” (p. 27). One major reason for a lack of subdomain distinctiveness is that testing programs are often retrofitting subscores from essentially unidimensional assessments that were not designed specifically to provide information at the subdomain-level. Attempting to report multidimensional score profiles from unidimensional tests, regardless of the psychometric model, applied will lead to a lack of subdomain distinctiveness (Luecht, Gierl, Tan & Huff, 2006). Therefore, if distinctiveness is lacking, decisions based on subdomain performance would be inaccurate.

The second concern of inadequate reliability is addressed in Standard 1.14, which maintains that “When a test provides more than one score...[the] reliability of the subscores should be demonstrated” (AERA, APA, & NCME, 2014, p. 27). Adequate subscore reliability is required to minimize errors in judgment when subdomain information influences decisions (Stone, Ye, Zhu, & Lane, 2010). As an example, some state educational testing systems utilize subscore performance to identify student learning needs and to plan educational interventions to meet these needs. However, if the subdomain information is not of adequate reliability such decisions can be based largely on measurement error, which from a practical context can lead to wasted resources in providing educational interventions that are not accurately directed toward an individual student’s learning needs.

To address the need for adequate reliability with subdomains that are often measured based on a small number of items, researchers have proposed using collateral

information (e.g., total score or performance on other subdomains) to improve subscore reliability (e.g., Wainer et al., 2001). Although these procedures have been shown to improve reliability, they do so at the cost of losing subdomain distinctiveness (Skorupski & Carvajal, 2010). As a result, there is still a need to find ways of reporting subscores that contribute unique diagnostic information and are statistically reliable.

1.4 Purpose of Study

Recent research has demonstrated that although subdomain information may provide no added value (i.e., distinct and reliable information of subdomain performance) beyond the total score, in some contexts such information is of utility to particular demographic subgroups (Sinharay & Haberman, 2014). Such a result suggests that when analyzing subscore added value for the total sample, subgroup differences are often masked. In most cases, this leads one to conclude that there is no subscore added value, which may lead to withholding diagnostic information as it is perceived to lack validity and adequate reliability for all examinees. However, in actuality, this information may be of particular use to identifying student learning-needs for a subgroup of examinees.

One limitation of previous research on evaluating comparability of subscore added value is that it has been evaluated for identifiable and protected demographic subgroups. However, it is argued that the utility of reporting subscores for an individual should not be based on one's manifest characteristics (e.g., gender or ethnicity), but rather on individual needs for diagnostic information, which is largely driven by a degree of multidimensionality in subdomain score profiles. However, when grouping examinees by manifest variables such individual multidimensionality can be masked if the majority of group members possess unidimensional data. As a result, individuals that would

benefit from diagnostic feedback would not be provided with such information due to their demographic membership.

To improve the validity of diagnostic information, this study proposes the use of multivariate outlier detection and non-parametric person-fit procedures to assess whether individual score profiles significantly depart from unidimensionality. Those examinees that are found to differ significantly can then be assessed separately for subscore added value. This approach has two major advantages over previous approaches. For one, it may serve as a way of reporting subscores that contribute unique diagnostic information and are statistically reliable. Secondly, it may avoid the perception that reporting differential subscore information for subgroups is discriminatory as within this approach groups are based on test performance as opposed to demographic membership. The importance of this study is clear in the wake of increased demand from stakeholders and the NCLB legislation for diagnostic information that accurately and reliably identifies student learning-needs.

Thus, the purpose of this study is three-fold and intends to answer, (1) How multidimensional do data need to be for subscores to have added value (i.e., be a better predictor of the true subscore than the observed total score)?; 2) How accurate are multivariate outlier detection and non-parametric person-fit statistics in identifying aberrant score profiles or response patterns due to multidimensionality?; and 3) When separating examinees into groups based on whether their score profiles or response patterns differ significantly from the total sample, does subscore added value invariance hold? The first question will address whether there are individuals that could benefit from diagnostic information as they possess multidimensional subscore domains, but are

masked by the largely unidimensional inter-subdomain correlations of the total sample. By demonstrating that there is an issue of masking effects when assessing subscore added value for some examinees, there will be justification for assessing aberrant score profiles or patterns due to multidimensionality, which is the focus of the second research question. The last question gets at a more general and important issue, which is when considering subgroups of examinees based on their score profiles or response patterns, do we obtain a different perception of subscore quality than when only considering the total sample?

These questions will be analyzed via simulation analyses, while the general approach of evaluating score profiles and response patterns will be applied to a large-scale applied dataset to evaluate its utility in practice. Results from this study are intended to: (a) further inform our understanding of subscore added value invariance when reporting raw subdomain scores and (b) provide an approach for reporting valid and reliable diagnostic information for those examinees who are of greatest need based on subscore profiles or response patterns rather than identifiable subgroup membership.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview of Literature Review

This chapter reviews the literature on subscore estimation procedures, how subscores are reported in practice, and validation approaches of subscore reporting. More specifically, this chapter can be outlined into the following five sections:

1. **Subscore Estimation Methodologies.** This section provides a review of existing methodologies that have been applied to subscore ability estimation. Methodologies will be divided by “simple” and augmented approaches. Within each approach, estimation procedures will be divided by classical test theory (CTT) and item response theory (IRT) frameworks.
2. **Comparative Analyses of Subscore Estimation Methodologies.** This section provides a review of studies that have evaluated the technical adequacy of subscore estimation methodologies by conducting comparative investigations. Within this section, comparative studies were divided based on the item type evaluated: a) dichotomous and b) ordinal.
3. **Subscore Reporting Practices.** This section summarizes reviews of how subscores are reported in practice by studying the literature related to score reports. Although this line of research is not focused primarily on the validity of subscore reporting, it provided useful guidance on the subscore estimation procedures that practitioners often employ.
4. **Validation of Operational Subscore Reporting.** This section reviews both methodologies for assessing the added value of reporting raw subdomain scores

and studies that have applied these methods to evaluating the validity of reporting raw subdomain scores as diagnostic information. As the focus of this paper is primarily on the use of fine-grained information to improve student learning, only studies assessing the validity of reporting individual-level subscores will be reviewed. Furthermore, as the comparability of scores across subgroups is important according to the *Standards* (AERA, APA, & NCME, 2014), studies that have evaluated subscore added value invariance will also be reviewed.

5. Summary Based on Literature Review. This section will summarize the findings from the four previous sections to provide justification for the need to conduct the current study.

2.2 Review of Subscore Estimation Methodologies

There are currently two general approaches to reporting subscores: 1) “simple” and 2) augmented procedures. The former approach estimates subscore ability by either calculating raw or percent-correct scores or by applying unidimensional item response theory (UIRT) estimation. For simple scores within the CTT framework, the added value of reporting subscores over total scores is assessed using two general methods: 1) Haberman’s (2008) method and 2) Brennan’s (2012) *Utility Index*. The simple IRT procedures include: 1) application of a unidimensional IRT model to items within each subtest separately, and 2) subscores based on a unidimensional model for each subtest, but with item parameter estimates based on the total number of items on the test. In contrast, augmentation approaches use collateral information (i.e., the total score or scores from other subdomains) to improve the stability of subscore estimation, which can be done using either CTT or IRT. The CTT procedures include: regressed estimates based

on univariate regression (Kelley, 1947) and Wainer et al.'s (2001) *subscore augmentation*. The remaining four procedures are generally based within the IRT framework: 1) Wainer et al.'s (2001) subscore augmentation applying IRT theta estimates, 2) Yen's (1987) *Objective Performance Index*, 3) the *Out-of-Scale Information* method (Kahraman & Kamata, 2004), and 4) subscores based on multidimensional IRT (de la Torre & Patz, 2005). However, one must note that any of the IRT methodologies could be further broken down into how subscores are reported (e.g., thetas, scale scores, IRT true scores, or percent-correct IRT true scores), as well as estimation methods and observed data types for theta estimates (Item Pattern: ML, MAP, and EAP; Summed Raw Score: ML, MAP, EAP, and raw to IRT scale score conversion).

2.2.1 Simple Subscore Estimation Methods

2.2.1.1 Number-Correct or Percent-Correct Raw Subscores

Of all the subscore estimation procedures, the number-correct or percent-correct raw subscores are the easiest to implement. This method sums the total number of correct responses on the subdomain of interest and either reports this raw score or calculates the total percent correct. This method is advantageous in that it can be computed quickly and it does not require advanced psychometric training to implement, which makes it intuitively comprehensible by non-technical audiences (e.g., school personnel and parents). However, as noted by Md Desa (2012), summed-scores have been judged to be unattractive for stakeholders when scores are subjected to public scrutiny in large-scale testing. Furthermore, one of the major disadvantages of this approach is that subscores do not necessarily accurately reflect actual strengths and weaknesses as examinees with the same raw scores are perceived as being of equal ability regardless of which items were

answered correctly; however, this is not a shortcoming of the methodology specifically, but is rather a characteristic of the CTT framework (Hambleton & Jones, 1993). An additional disadvantage of this approach is that subscore reliability is not considered or augmented as the other CTT methods later described in this section, which is particularly troubling as the reliability of subscores is one of the major difficulties to valid subscore reporting.

2.2.1.2 Simple Unidimensional IRT Subscore Estimation Procedures

One approach that is common in large-scale assessments is to estimate subscores by assuming an independent unidimensional space for each subtest (Buluth, 2013). More specifically, items within a subdomain are calibrated separately to obtain theta estimates for that subdomain separately. In practice, this would require $n+1$ calibrations, where n is equal to the number of subdomains and the additional calibration would be that of the overall score where all items would be included. Another approach is a two-stage process for obtaining subdomain ability estimates. Within this approach, a unidimensional IRT model is first applied to all items within a test. Next, subdomain ability estimates are obtained via fixed item parameter estimation based only on those items that belong to the targeted subdomain. For an applied example of this approach, the reader is referred to Bock, Thissen, and Zimowski (1997).

2.2.2 Subscore Augmentation Estimation Methodologies

2.2.2.1 Kelley's (1947) Univariate Regression

Kelley's regressed-score estimates (RSEs) are based on the linear regression of true score (T) on observed score (X), which results in the following equation:

$$\tilde{T} = \rho_X^2 X + (1 - \rho_X^2)\bar{T}, \quad (1)$$

where \tilde{T} is the estimated true-score random variable based on the linear regression, ρ_X^2 is the reliability of the observed score, and \bar{T} is the mean true score. Under CTT assumptions, $\bar{T} = \bar{X}$, which allows for equation 1 to be expressed as:

$$\tilde{T} = \rho_X^2 X + (1 - \rho_X^2) \bar{X}, \quad (2)$$

where \bar{X} is the observed score mean. As noted by Tao (2009), equation 2 has an empirical Bayesian interpretation as the essence of this approach is to remove the unreliable part of the observed score by regressing it to the mean. Specifically, as the reliability of the observed score increases, \tilde{T} is more influenced by X . However, as the reliability of the observed score decreases, \tilde{T} is increasingly influenced by \bar{X} . As an extreme, if ρ_X^2 is equal to 0, \tilde{T} is equal to \bar{X} ; however, if ρ_X^2 is equal to 1, \tilde{T} is equal to X , which would mean that every examinee's RSE is equal to his/her observed score. As a result, Kelley's method utilizes the observed score mean as collateral information to account for unreliable subscores.

2.2.2.2 Wainer et al.'s (2001) Subscore Augmentation

Wainer et al. (2001) extended Kelley's (1947) method by using regressed-score estimates that are based not on the observed score mean, but rather are based on information from other subscores in a multivariate context. Equation 2 can be algebraically rearranged to the following equation:

$$\tilde{T} = \bar{X} + \rho_X^2 (X - \bar{X}) \quad (3)$$

and can be represented in a multivariate context as:

$$\tilde{T} = \mathbf{X} + \mathbf{B}(\mathbf{X} - \mathbf{X}.), \quad (4)$$

where \mathbf{X} is a vector of subdomain means, \mathbf{X} is a vector of subdomain scores, and \mathbf{B} is a matrix that is the multivariate analog for the estimated reliability and is estimated as follows:

$$\mathbf{B} = \Sigma_T \Sigma_X^{-1}, \quad (5)$$

where Σ_T is the true score covariance matrix and Σ_X^{-1} is the inverse of the observed score variance and can be obtained directly from the sample. This solution is based on the CTT definition of reliability as the proportion of true score variance relative to the observed score variance. Under the CTT assumption, true scores are uncorrelated with error, which means that the off-diagonal elements within both Σ_T and Σ_X will be equal. Therefore, to obtain the diagonal elements of Σ_T , one must obtain the diagonal elements of Σ_X by the reliability (coefficient α) of the subdomain of interest. Therefore, the empirical Bayes estimate of the vector of true subscores, τ_p , for examinee p , conditioned on observed scores, is:

$$E(\tau_p | x_p) = \mu + \Sigma_T \Sigma_X^{-1} (x_p - \mu), \quad (6)$$

which is estimated in practice as:

$$\tilde{T} = \mathbf{X} + \Sigma_T \Sigma_X^{-1} (\mathbf{X} - \mathbf{X}.) = \mathbf{X} + \mathbf{B}(\mathbf{X} - \mathbf{X}.) \quad (7)$$

An estimate of the conditional covariance matrix of the estimated true score can also be obtained to compute the conditional standard errors for augmented subscores (See Wainer et al., 2001).

As noted by Tao (2009), Wainer et al.'s (2001) and Kelley's (1947) methods are identical when the off-diagonal elements of \mathbf{B} are equal to zero, which would indicate that the subscores are independent. As a result, when the subscores are perfectly reliable, the estimated true score for an examinee is equal to his/her observed score, while all

observed scores are regressed to the mean when reliability is equal to zero. The differences between the two methods are apparent when the off-diagonal elements are not equal to zero. When this is the case, Wainer et al.'s method allows for borrowing information from other subscores to augment the reliability of the subscore of interest.

2.2.2.3 Wainer et al.'s (2001) Augmentation Method with IRT Theta Estimates

As noted by Wainer et al. (2001), testing programs may prefer reporting IRT scale scores as opposed to number or percent-correct scores, which requires the need to develop augmentation procedures that can generalize the empirical Bayes approach for application with IRT scale scores. To do this, the authors adapted the CTT approach described in section 2.2.2.2. Specifically, this procedure requires unidimensional IRT ability estimates obtained using maximum likelihood (MLE), maximum a posteriori (MAP), or expected a posteriori (EAP) methods. If MAP or EAP estimates are applied, there is first a need to correct these ability estimates due to their tendency to shrink to the population mean (Fu & Qu, 2012). Assuming that the population mean is 0 and the standard errors are constant, the correction² is made as follows:

$$\text{MAP}^*(\theta_s) = \frac{\text{MAP}(\theta_s)}{\rho_s}, \quad (8)$$

where $\text{MAP}^*[\theta_s]$ is the corrected IRT scale estimate on subscale s , $\text{MAP}(\theta_s)$ is the original estimate obtained from the unidimensional calibrations, and ρ_s is an estimate of the reliability of subscale s , which is calculated as:

$$\rho_s = \frac{\sigma^2 \text{MAP}(\theta_s)}{\sigma^2 \text{MAP}(\theta_s) + \sigma_e^2}, \quad (9)$$

² This correction can be applied to either MAP or EAP theta estimates. For simplicity's sake, only the MAP correction is illustrated.

where $\sigma^2 \text{MAP}(\theta_s)$ is the variance of the IRT scale scores for subscale s and $\bar{\sigma}_e^2$ is the average value of the variances of the error of measurement associated with those scores. These corrected theta estimates are then applied to equation 4 and augmentation is conducted in the same way as with observed scores. It should be noted that if MLE is used to obtain ability estimates, there is no need to apply the correction procedure described in this section (Fu & Qu, 2012).

2.2.2.4 Objective Performance Index

To improve the stability in reporting subscores, Yen (1987) proposed the Objective Performance Index (OPI), which is a procedure that combines subdomain performance with information from the examinee's overall test performance to provide stability in reporting subscores. More specifically, it implements a Bayesian IRT estimation to obtain an estimated true score (estimated proportion-correct) for items on a subdomain given their overall test performance. This is accomplished in five steps. First, item parameters are estimated for the entire test using an IRT model, such as the three-parameter logistic (3PL) model:

$$P_i(\theta_j) = c_i + (1 - c_i) \frac{\exp[1.7a_i(\theta_j - b_i)]}{1 + \exp[1.7a_i(\theta_j - b_i)]} \quad (10)$$

where $P_i(\theta_j)$ is the probability of correctly answering item i given examinee j 's ability, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the pseudo-guessing parameter. Secondly, ability estimates ($\hat{\theta}$) are obtained for each examinee by treating the item parameter estimates (a_i, b_i, c_i) as fixed. Upon obtaining item and ability estimates, a true score for each examinee is estimated for performance on the targeted subdomain by plugging parameter estimates into equation 10 for those items within the targeted subdomain:

$$\hat{T}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\hat{\theta}), \quad (11)$$

where \hat{T}_j is the true score (expected proportion correct) for subdomain j , n_j is the number of items in subdomain j , and $P_{ij}(\hat{\theta})$ is the probability of correctly responding to item i in subdomain j . The fourth step is to determine whether the estimated true score for an examinee is consistent to what would be expected. That is, one issue with computing an expected proportion correct is that for some examinees a subdomain may be multidimensional. For example, an examinee may be familiar with international history, but may have little knowledge of domestic history due to their immigrant status. Therefore, the author developed the following statistic to evaluate unexpected subdomain performance given the examinee's observed percent-correct subscore:

$$Q = \sum_{j=1}^J \frac{n_j \left(\frac{x_j}{n_j} - T_j \right)^2}{\hat{T}_j (1 - \hat{T}_j)}, \quad (12)$$

where J is the total number of subdomains. The Q statistic is then compared to the critical value from a chi-square distribution with J degrees of freedom at an alpha-level of .10. The last step of the analysis is to compute OPI; however, the Q statistic impacts how the OPI is calculated. For example, if $Q \leq \chi^2(J, 0.10)$, the OPI (\tilde{T}_j) is defined as a weighted average of the observed subscore and the estimated subscore:

$$\tilde{T}_j = \frac{\hat{T}_j n_j^* + x_j}{n_j^* + n_j}, \quad (13)$$

where

$$n_j^* = \frac{\mu(\hat{T}_j | \theta) [1 - \mu(\hat{T}_j | \theta)]}{\sigma^2(\hat{T}_j | \theta)}, \quad (14)$$

$$\mu(\hat{T}_j | \theta) \approx \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\hat{\theta}), \quad (15)$$

$$\sigma^2(\hat{T}_j|\theta) \approx \frac{\left[\frac{1}{n_j} \sum_{i=1}^{n_j} P'_{ij}(\theta)\right]^2}{I(\theta, \hat{\theta})}, \quad (16)$$

where

$$P'_{ij}(\theta) = \frac{1.7a_{ij}[1-P_{ij}(\theta)][P_{ij}(\theta)-c_{ij}]}{1-c_{ij}}, \quad (18)$$

and if theta is estimated using the maximum likelihood estimation procedure based on the examinee's number correct score,

$$I(\theta, \hat{\theta}) = \frac{\left[\sum_{j=1}^J \sum_{i=1}^{n_j} P'_{ij}(\theta)\right]^2}{\sum_{j=1}^J \sum_{i=1}^{n_j} [P_{ij}(\theta)[1-P_{ij}(\theta)]}. \quad (19)$$

If there are items that do not contribute to any subdomain but do participate in the estimation of ability, the information contributed would be added to equation 19. If $Q > \chi^2(J, 0.10)$, the OPI (\tilde{T}_j) disregards prior information and is defined as the observed percent-correct subscore:

$$\tilde{T}_j = \frac{x_j}{n_j}. \quad (20)$$

2.2.2.5 Out-of-Scale Information Method

Building upon the work of Davey and Hirsh (1991) and Ackerman and Davey (1991), Kahraman and Kamata (2004) proposed the Out-of-Scale method. Within this procedure, subscore ability estimation is augmented by using collateral information from other subdomains. More specifically, this procedure can be conceptualized in three major steps. First, item parameters for items of the subdomain of interest (“in-scale” items) are first calibrated using for example, the 3PL IRT model. Secondly, each item outside of the targeted subdomain (“out-of-scale” items) are calibrated using maximum likelihood estimation by holding constant the in-scale item parameters for each targeted subdomain. Holding constant the in-scale item parameters is meant to ensure that the individual out-

of-scale items are calibrated with respect to the domain trait that is defined in the in-scale items. This estimation procedure can be expressed as follows:

$$L(\mathbf{U}|\boldsymbol{\theta}, a_{n+1}, b_{n+1}, c_{n+1}) = \prod_{j=1}^s \prod_{i=1}^n P_i(\theta_j)^{u_{ij}} [1 - P_i(\theta_j)]^{1-u_{ij}} \times P_{n+1}(\theta_j)^{u_{(n+1)j}} [1 - P_i(\theta_j)]^{1-u_{(n+1)j}}, \quad (21)$$

where \mathbf{U} is the response matrix with all in-scale items and one out-of-scale item, $\boldsymbol{\theta}$ is the ability parameter vector for all examinees, u_{ij} is the item response for item i , person j , P_i is the item response function for item i given the known parameters from the in-scale items obtained from the 3PL model, P_{n+1} is the item response function for item $n + 1$ (i.e., the one out-of-scale item being calibrated) with unknown item parameters $(a_{n+1}, b_{n+1}, c_{n+1})$ using the 3PL model. The last step is to estimate examinee ability on the targeted subdomain by using the expected a posteriori (EAP) estimator for item responses on *both* the in-scale and out-of-scale items.

2.2.2.6 Multidimensional IRT Applications to Subscore Estimation

The IRT subscore estimation procedures that have been described up to this point are based on unidimensional modeling of the data. However, as noted by Buluth (2013), these methodologies are limited in a number of ways. For one, a simple structure (i.e., each item is only an indicator for one subdomain) is assumed and secondly, the covariance among the subdomains is ignored. The impact of the latter limitation on unidimensional theta estimation when the underlying structure of the data is multidimensional was examined by Tate (2004) who found that the standard errors of the unidimensional ability estimates increased as the number of dimensions increased and the covariances among the latent dimensions decreased. To account for possible biasing in

subscore estimation when applying unidimensional procedures to data that are multidimensional in nature (i.e., reporting subdomains automatically assumes a multidimensional data representation), researchers have suggested the application of multidimensional IRT (MIRT) models.

MIRT models differ from unidimensional IRT (UIRT) models in multiple ways. For one, MIRT models extend UIRT models by modeling two or more latent dimensions simultaneously. Furthermore, MIRT models can estimate both simple and complex test structures (i.e., an item can be an indicator of more than one latent dimension). In terms of complex test structures, the probability of correctly responding to an item is based on a vector of ability as opposed to a single ability. Complex MIRT models can either be compensatory or non-compensatory.

2.2.2.6.1 Compensatory MIRT Models

Compensatory models allow for a high ability on one dimension to offset or compensate for a low ability on another dimension. For example, in a math word problem, if the examinee possessed low reading ability, but high math ability, the probability of a correct response would be moderate. Allowing one dimension to offset another dimension is clearly reflected in the summation of the logit in the multidimensional two-parameter logistic model (M2PLM):

$$P(x_i = 1) = \frac{e^{\sum_{k=1}^m a_{ik}(\theta_k + A_i \Delta_i)}}{1 + e^{\sum_{k=1}^m a_{ik}(\theta_k + A_i \Delta_i)}}, \quad (22)$$

where m is the number of dimensions, \mathbf{a}_{ik} is a vector of k slope estimates for item i , θ_k is a vector of k ability estimates, A_i is the multidimensional discrimination, and Δ_i is the multidimensional item difficulty. A number of compensatory MIRT models have been proposed for estimating subscores. These models include, but are not limited to,

multidimensional extensions of the 1PL, 2PL, 3PL, partial-credit, and graded-partial credit model (e.g., Adams, Wilson, & Wang, 1997; Béguin & Glas, 2001; Haberman, von Davier, & Lee, 2008; McDonald, 1997; Muraki & Carlson, 1995; Reckase, 1997; von Davier, 2008). Furthermore, as the most basic models estimated using item factor analysis can be viewed as reparameterized extensions of the 2PL model, the types of compensatory multidimensional models that can be estimated are quite vast (e.g., the higher-order and bifactor models; de la Torre & Song, 2009; Md Desa, 2012).

2.2.2.6.2 Non-Compensatory MIRT Models

In contrast to compensatory models, ability on one dimension does not compensate for ability on the other dimension(s) (de Ayala, 2009). Researchers have argued that the compensatory modeling approach is unrealistic to actual cognitive processes that occur when solving a test item (Ackerman, 1989). For example, if one has low reading ability, his/her probability of correctly responding to a math word problem should be low regardless of math ability as reading comprehension is required to solve the problem. Such an outcome is modeled by the multiplicative nature of the logit as demonstrated in the non-compensatory 2PLM:

$$P(x_i = 1) = \prod_{k=1}^m \frac{e^{a_{ik}(\theta_k + A_i \Delta_i)}}{1 + e^{a_{ik}(\theta_k + A_i \Delta_i)}}, \quad (23)$$

which is essentially the product of 2PL models for m dimensions. Although non-compensatory models seem to provide more realistic modeling as a deficit on one dimension cannot be overcome by a strength on other dimensions, compensatory models are still more popular. As an example, non-compensatory models that provide continuous latent variables are limited to Sympson (1978), Whitley (1980), and Embretson (1984).

2.2.2.6.3 Cognitive Diagnostic Models

CDMs have seen an increase in popularity due to their purported ability to provide more formative feedback than common IRT models. At the most basic level, CDMs differ in that they present examinees with information concerning mastery of discretely defined (mastered or unmastered) skills or abilities, whereas traditional IRT models provide an estimate of ability that is on a continuous scale. As noted by Lee and Sawaki (2009), the procedure for using CDMs to provide diagnostic feedback is as follows: (a) identify overall skills that are measured by a task (i.e., an item), (b) list the skills that are required for successfully answering an item (this is done for all items on the test), (c) apply a CDM to estimate the profiles of skill mastery for an examinee based on test performance, and (d) disseminate diagnostic feedback to stakeholders. This process joins both cognitive science and psychometrics to make assumptions about (a) the cognitive processes (skills) that are required by an examinee to complete a task and (b) the item characteristics that are intended to elicit these cognitive processes (Jang, 2008).

According to Fu and Li (2007), at least 62 CDMs have been proposed in the literature. Although these models clearly differ to some degree, they all share a number of similar characteristics. For one, all models provide diagnostic information via multidimensional confirmatory modeling. The confirmatory nature of CDMs comes from the substantive definition (hypothesis) of the multiple skills (multidimensionality) necessary to complete tasks on the test, which is specified in the Q matrix. Furthermore, all CDMs allow for multiple criterion-referenced interpretations as the most basic CDMs provide a single cut-value on the latent dimensions separating mastery and non-mastery of skills. One of the last similarities is that most applications of CDMs consist of complex

factor loadings (Rupp & Templin, 2008); however, this leads to differences in CDMs on whether data are modeled in a compensatory or non-compensatory manner. Additional differences between CDMs lie in the types of observed response variables that the models can handle, as well as the scale of the latent variables (dichotomous or polytomous). Although CDMs provide both flexibility in modeling and have the potential to provide diagnostic information, they are rarely applied in current large scale assessments due to a lack of fine-grained understanding of the cognitive processes underlying many skills (Fu & Qu, 2012).

Overall, MIRT allows for flexible modeling of subdomain performance as compensatory, non-compensatory, or latent class models can be specified. Furthermore, MIRT may provide a more straightforward approach to subscore estimation when compared to empirical Bayes augmentation procedures, such as the OPI and subscore augmentation methods, that require multiple steps (Buluth, 2013). For such methods, unidimensional parameter calibration is first conducted and then ancillary information is used to improve the precision of subscores. In contrast, MIRT models obtain ancillary information, such as subdomain covariances, within a single estimation procedure. That is, although parameter estimation of MIRT models is more complicated than UIRT models, it is, to a certain extent, more efficient (Fu & Qu, 2012).

2.3 Comparative Analyses of Subscore Estimation Methodologies

2.3.1 Comparison of Methods for Dichotomous Items

As there are a number of subscore estimation procedures that have been proposed in the literature, researchers have conducted comparative studies to provide recommendations on which procedures are most appropriate for practical use. As an

example, Bock, Thissen, and Zimowski (1997) sampled and resampled subsets of operational data (15-20 items of 100 total items) to compare the number correct, the maximum likelihood IRT (ML-IRT) percent-correct, and the Bayes IRT percent-correct subscore estimation procedures. As the number correct of the 100 items was known, root mean squared errors (RMSEs) of the predicted domain scores were computed. Results demonstrated that for the 15-item samples the average RMSE across 30 replications was 10% and 50% smaller for the ML-IRT and Bayes percent-correct procedure, respectively, when compared to the CTT number correct procedure. The same RMSE differences between the ML-IRT percent-correct and the number correct procedure were observed under the 20 item condition, while the RMSEs for the Bayes percent-correct procedure were found to be the most superior (70% smaller than number-correct). Taken together, these results suggest that the IRT estimator was a more accurate predictor of the domain score than the CTT number correct method.

Luecht (2003) compared four approaches: 1) standardized number correct scores (ZX), 2) EAP scores based on a unidimensional total-test 3PL model calibration (UIRT-T), 3) MAP scores based on separate unidimensional 3PL model calibrations for the separate subdomains (UIRT-S), and 4) MAP scores based on a multidimensional 3PL model calibration (MIRT). Data were generated for 74 items that were modeled using a four-factor, oblique simple-structure MIRT model for 2,000 simulees. Dependent variables examined included: 1) subdomain correlations, 2) measurement errors, and 3) diagnostic score profiles. Results demonstrated subdomain correlations among the ZX, UIRT-T, and UIRT-S to all be near 1.0. In terms of standard errors, the UIRT-T subscores were found to have the largest standard errors (ranged from 0.43 to 1.61),

while the ZX procedure produced standard errors ranging from 0.42 to 0.78 across subdomains. Both UIRT-S and MIRT produced similar standard errors. The UIRT-T and UIRT-S were found to produce score profiles that were most similar to the true profiles; however, overall, subscore profiles were found to differ greatly, which points to the fact that the choice of subscore estimation procedure can greatly impact remediation decisions.

Edwards and Vereza (2006) compared Wainer et al.'s (2001) subscore augmentation method with IRT EAP estimates and raw subscores. Data were simulated based on a 3PL model for tests that differed in the number of subdomains, the number of items within a subdomain (reliability), and subscale correlations, which resulted in a total of 30 conditions. Specifically, simulated tests consisted of either two or four subdomains with subdomain correlations being equal to 0.3, 0.6, and 0.9. Four subscale lengths were chosen to simulate subscores that were either unreliable ($\alpha = .43$ or $\alpha = .59$) or reliable ($\alpha = .75$ or $\alpha = .85$). Across all conditions, the sample size was constrained to 2,000. The comparison of augmented versus non-augmented scores was compared in terms of RMSE (square root of the average squared difference between estimated and generated thetas), reliability (square of the correlation between true and estimated thetas), the percentage of simulees that had estimated augmented scores closer to truth than non-augmented scores, and classification accuracy.

Results of the analysis were presented only for the two subdomain conditions as the findings were very similar to the four subdomain conditions. RMSE values were found to be relatively similar between the two subscore estimation procedures when both subdomains possessed either 10 or 20 items and a correlation of 0.6. In such conditions,

the reduction in RMSE for Wainer et al.'s (2001) method was found to be equal to 5%. However, in more extreme conditions, such as when the subdomains differed greatly in reliability and the subdomain correlation was equal to 0.9, RMSE reduction for the augmentation method was found to be equal to 33%. More realistic conditions of equal sample sizes and high correlations (0.9), demonstrated that when the number of subdomain items was equal to 5 or 10, RMSE differences ranged from 0.10 to 0.15. As expected, when there were a large number of subdomain items (20 or 40), RMSE differences decreased as the collateral information was not useful in improving the already reliable subdomains. In terms of reliability, the augmented procedure provided greatest improvements when the subdomain providing collateral information was much more reliable than the targeted subdomain by as much as 1.5 times. However, under conditions where the number of subdomain items was equal and the subdomain reliabilities was 0.3 or 0.6, the improvements in augmented reliability decreased between 0.2 to 0.3. In examining the similarity between estimated and true thetas, the augmented procedure was found to more accurately estimate ability for simulees across all conditions by an average of 5% (ranged from 1% to 16%). Augmented scores were also found to improve classifications by 0.03% to 13.43% depending on the condition. Under realistic conditions (equal subdomain length and high inter-subdomain correlations), classification accuracy improved by only 0.15% to 3.82%. Overall, the results of this analysis demonstrated that the improvements gained from using an augmented procedure are a function of subdomain correlations and subdomain reliability.

2.3.2 Comparison of Methods for Mixed Format Tests

The two previous studies examined the accuracy of subscore estimation methods only in the context of dichotomous tests. To fill the gap in the literature, Shin (2007) compared the following methods for mixed-format tests: 1) percent-correct raw subscores, 2) 3PL/GPCM IRT true subscore based on item parameters estimated on in-scale items only, 3) OPI with 3PL/GPCM theta estimates based on all items, 4) Wainer et al.'s (2001) score augmentation based on raw scores, and 5) Wainer et al.'s (2001) method with MCMC theta estimates. Four simulation factors were included: 1) number of examinees (250, 500, and 1,000), 2) test length (6, 12, or 18 items per subdomain), 3) subdomain correlations (0.5, 0.8, and 1.0), and ratio of constructed response (CR) over multiple-choice (MC) items (0%, 20%, and 50%), which resulted in 81 conditions. The dependent variables examined were objective score reliability, bias, and RMSE.

Reliability was found to be impacted minimally by sample size for all methods, except for method 2. Furthermore, differences in reliability were less than 0.1 across the different test lengths for all procedures. Interestingly, the proportion-correct method was found to have higher reliability than method 2 when the test length was equal to 18. When subdomain correlations were equal to 0.8, the differences in reliability were as great as 0.05. The largest difference in reliability for the proportion-correct method when compared to the other methods occurred when the subdomain correlations were equal to 1.0. In terms of bias, independent variables including test length, subdomain correlations and the ratio of CR to MC items had the largest impact. Across all of these conditions, bias was found to be smallest for the proportion-correct method, while the magnitude differences for test length, subdomain correlations, and the ratio of CR to MC items was

equal to 0.01 to 0.05, 0.01 to 0.09, and 0.01 to 0.08, respectively. RMSE differences across conditions were found to be minimal for all methods (0.01 to 0.05). Overall, using augmentation methods maximally improved reliability by about 0.09 points and RMSE by approximately 0.05 points when compared to the non-augmented method.

An additional study that examined subscore estimation methods in mixed-format tests was conducted by Yao and Boughton (2007). Six methods were included: 1) percentage correct on subscale number-correct scores (NC), 2) multidimensional IRT Bayesian subscale scores (BMIRTSS), 3) multidimensional IRT Bayesian domain subscale scores (BMIRTDS), 4) OPI subscale scores, 5) an IRT pattern subscale scoring approach using maximum likelihood estimation (MIRTPSS), and 6) a unidimensional IRT objective-level Bayesian scoring approach (UIRTOJSS). Data were generated for a four-dimension simple structure model consisting of a total of 60 items and subdomain lengths ranging from 12 to 18 items. The accuracy of subscore estimates and classification accuracy were compared when varying sample size (1,000, 3,000, 6,000) and subdomain correlations (0, 0.1, 0.3, 0.5, 0.7, 0.9). As BMIRTSS, MIRTPSS, and UIRTOJSS were expressed on a three-digit latent score metric they were compared amongst each other, while the remaining procedures were compared separately.

Results demonstrated that across all conditions, BMIRTSS provided improved ability estimate recovery when compared to MIRTPSS. Additionally, BMIRTSS and UIRTOJSS were found to have similar recovery when correlations were low (0.1); however, as the correlations increased, BMIRTSS provided improved recovery, while UIRTOJSS provided similar recovery across all conditions largely as it did not use the subdomain correlations in ability estimation. RMSE values were found to be smaller

across all conditions for BMIRTDS when compared to the OPI and NC methods. The largest differences were observed when the subdomain correlations increased as BMIRTDS utilized this information for estimation, while the OPI did not. Furthermore, sample sizes had little impact on the results as 1,000 simulees were sufficient. In terms of classification accuracy, NC was found to have the largest errors across all conditions with approximately 65% misclassification. In contrast, the lowest rates were observed for MIRTSS. As the correlations increased, BMIRTSS and BMIRTDS provided classification errors at similar rates to the OPI. Overall, results of this study demonstrated the utility of applying MIRT models for subscore estimation and classification accuracy when compared to CTT augmented and non-augmented procedures.

One study that solely compared augmented estimation procedures was conducted by de la Torre, Song, and Hong (2011). In this study, the augmented estimation procedures compared were: 1) Wainer et al.'s (2001) augmentation method (AS), 2) the higher-order item response model using MCMC estimation, 3) Bayesian multidimensional scoring using MCMC estimation and 4) the OPI. Data were generated based on the higher-order item response model for a fixed sample size of 1,000 simulees. The independent variables manipulated included the number of subdomains (2 and 5), test length (10, 20, and 30), and subdomain correlations (0, 0.4, 0.7, and 0.9), which resulted in a total of 24 conditions. The dependent variable of interest was ability parameter recovery, which was examined in terms of correlations, RMSE, bias, and estimated proportion correct.

Correlation analyses demonstrated that the OPI systematically underestimated ability, while the other procedures differed in correlations minimally (0.00 to 0.10). The

largest differences were observed when there were five subdomains, subdomain correlations of 0.90, and test lengths of 10. RMSE results suggested that as the test length increased, RMSE decreased. Such a pattern was consistent across methods and conditions. In terms of conditional bias, the greatest differences of ability estimates were seen at the extremes of the theta continuum (-1.75 to 1.75), whereas the majority of procedures were in high agreement around average ability. This result is reflected in differences in estimated proportion correct where across all conditions the differences ranged from 0 to 0.01. Overall, the results from this study suggested that the subscore estimation methodologies provided very similar results. The largest differences were observed at the extremes of the theta continuum where the higher-order and multidimensional scoring procedures provided more accurate results.

An additional study that solely compared the utility of augmented subscores was conducted by Skorupski and Carvajal (2001). Whereas, Torre, Song, and Hong (2011) primarily investigated the advantages of multidimensional IRT approaches, the authors for this study were interested in unidimensional IRT approaches. More specifically, the three methods analyzed within this study included: 1) the OPI, 2) Wainer et al.'s (2001) method with raw scores, and 3) Wainer et al.'s (2001) method with IRT ability scores. Comparative analyses were based on real data that came from a statewide testing program consisting of four subdomains (subdomain length ranged from 11 to 15 items) for 17,266 examinees. The dependent variables examined included: 1) average change in examinee subscore ability estimates, 2) change in subscore reliability, and 3) subdomain correlations before and after augmentation. Within this study, change was defined as the

difference in criteria between augmented and raw scores (either CTT or IRT depending on augmentation method).

In comparing raw CTT scores with method 2, the average sample mean subscores were identical. However, the augmentation method did reduce the standard deviations, which led to average squared change ranging from 1.38 to 1.82 for the four subdomains. Similar patterns were observed when comparing IRT raw scores with methods 1 and 3. As an example, the average subscores between methods differed by 0.01 to 0.04 points, while the average standard deviations were reduced by an average of 14% to 28%. However, the average squared changes were smaller than the CTT methods as they ranged from 0.36 to 0.78. In terms of reliability, the subdomain internal consistency reliabilities (α) improved from 15% to 30%. As expected, the largest reliability improved was provided for the subdomain that had the fewest items. Furthermore, reliability improvements were consistent across all three augmentation methods. However, such improvements in reliability came at cost. That is, after applying augmentation procedures the subdomain correlations ranged from 0.97 to 1.00 across methods 1 through 3. In contrast, the subdomain correlations of the original CTT raw scores ranged from 0.62 to 0.72. Overall, this study demonstrated the utility of using augmented scores for significantly improving subscore reliability, particularly when the subdomain test length was relatively short. However, this was accomplished in different ways by the augmentation methods. Specifically, the regression approaches (methods 2 and 3) increased reliability by making every examinee's score profile look more like the overall score profile, while the OPI method increased reliability by making all subscore means and standard deviations essentially the same across the subdomains. These findings

suggested that the cost of increasing the reliability of subdomain scores is the loss of diagnostic score meaning at the individual examinee-level. Therefore, we are left with the question of how to improve the reliability of subscores without exaggerating subdomain interrelationships.

2.4 Subscore Reporting in Practice

The previous sections of the literature review discussed different methods for estimating subscores and comparative studies that evaluated the technical adequacy of these procedures. This section will focus on two aspects: 1) how subscores are estimated and reported to stakeholders in practice and 2) reviewing previous validation studies that evaluated subscore added value to better understand the measurement characteristics (i.e., sample size, strength of subdomain correlations, and subdomain test length) that are required to support valid subscore reporting.

To evaluate how subscores are estimated in practice this section will rely on the literature related to feedback accessibility. One of the first and most well-known studies in this area was conducted by Goodman and Hambleton (2004). In this study, the authors sampled student score reports from 14 states (Connecticut, Delaware, Louisiana, Massachusetts, Minnesota, Missouri, New Jersey, Pennsylvania, Virginia, Wisconsin, and Wyoming), two Canadian provinces (British Columbia and Ontario), and three U.S. commercial testing companies (Harcourt Educational Measurement, CTB/McGraw-Hill, and Riverside Publishing). Although the authors focused on numerous aspects of score reporting, this study will rely primarily on their analysis of providing examinee-level diagnostic information. Within this study, diagnostic information was operationally defined as information that provided detail beyond the general subject-level.

Across 11 states, one province, and all three testing companies, diagnostic information was supplied to stakeholders in two ways: 1) results by subdomain and 2) specific skills that an examinee demonstrated or needs to improve. Furthermore, one Canadian province provided diagnostic information only to examinees that did not pass the respective subdomain. For these testing programs, subdomain results were reported numerically as raw scores, percent correct scores, or percentile rank scores. In terms of precision, no states or provinces provided reliability estimates of subdomain scores and only two commercial test publishers depicted confidence intervals when reporting subscores. In addition to reporting numerical subdomain performance, two states and one province reported particular strengths and weaknesses of individual students on the respective subdomain. These results led the authors to recommend that when reporting subdomain performance testing programs should report only scale scores, as well as validity and reliability evidence. Furthermore, they recommended that more testing programs should include customized interpretations of examinee subdomain performance, which would include concrete and easily-implemented suggestions to improve future performance.

A more recent study was conducted by Wang, Faulkner-Bond, and Shin (2012) in which subscore reporting practices for K-12 English language arts (ELA) assessments were evaluated across 46 states in the U.S. Score reports were coded for a number of important features including: (a) the presence of diagnostic information and the technique of reporting such information, (b) the types of subscores (i.e., raw scores, scale scores, performance-level descriptors), and (c) subscore reliability. Overall, 41 states were found to report subscore information that was most often reported using raw scores (28 states);

however, 15 states were found to combine raw and percent correct scores. Additionally, five states were found to report only scale scores, while six combined raw, percent-correct, and scale scores. Interestingly, three states that only reported scale scores based such information on IRT expected scores. Four states were found to report either performance-level or performance-level descriptors without reporting any numerical scores, while 12 other states combined such descriptors with numerical scores.

In terms of reporting subscore precision, only six states provided confidence intervals; however, nine states included some cautioning that the reported subscores could be of low reliability. Overall, the authors found that states use a variety of approaches to reporting subscores. Although the most popular approach was to report raw subdomain scores, some states attempted to convey subdomain performance in creative ways, such as through performance levels, norm-referenced scores, and projected scale scores. These creative approaches were most likely driven as only 11 states reported one or more subscores that had a test length of 20 or more items. As a result, the authors suggest that further improvements are required in both test development and subscore reporting practices to provide more useful diagnostic information for improving student performance.

Faulkner-Bond et al. (2013) extended the work of the previous two studies by specifically investigating score reporting practices on English language proficiency (ELP) assessments. In total, ELP score reports were evaluated for 24 individual states and one consortium (included 26 states and the District of Columbia). The authors found that across all testing programs subscores were reported and one state provided “next steps” for improving examinee ELP by subdomain performance. Subdomain performance for

each testing program was reported using scale scores, while five states additionally reported subdomain performance based on raw scores. According to the authors, the additional reporting of raw scores may have been due to a general misconception of scale score meaning by stakeholders (Trout & Hyde, 2006). Of the 27 score reports reviewed, only two provided measurement error or precision related to subdomain performance. Furthermore, only one of these states actually reported the meaning of measurement precision. Based on these findings, the authors recommended that testing programs should report the precision of subdomain performance as well as ensure reliability and utility. More specifically, they suggested that if subscores are not precise or reliable enough to provide added value, such information should not be reported.

2.5 Validation of Subscore Reporting in Practice

As the previous section has highlighted that the majority of testing programs for K-12 content and ELP assessments report raw subscores as diagnostic information, this section will review: 1) methods for assessing the added value of reporting raw subscores, and 2) research that has evaluated validity evidence for reporting subscores based on raw or percent-correct scores. The latter objective will focus specifically on validity studies based on individual-level and group-level invariance analyses. From this review, recommendations will be made regarding measurement characteristics (e.g., sample size, subdomain test length, and subdomain inter-correlations) that are necessary for obtaining subscore added value.

2.5.1 Methods for Assessing the Adequacy of Reporting Raw Subscores

2.5.1.1 Haberman's (2008) Method

Haberman's (2008) procedure evaluates whether subscores provided added value by assessing whether the observed subscore is a better predictor of the true subscore when compared to the observed total score. Within this framework, observed subscore and observed total score predictors of the true subscore estimates are as follows, respectively (Sinharay, 2010; Sinharay, Puhan, & Haberman, 2011):

$$s_s = \bar{s}_s + \alpha(s - \bar{s}_s), \quad (24)$$

where s_s = the true subscore estimate based on the observed subscore, \bar{s}_s = the observed mean subscore for the sample, α is equal to the reliability of the subscore, s = the observed subscore for subtest s , and

$$s_x = \bar{s}_s + c(x - \bar{x}), \quad (25)$$

where s_x = the true subscore estimate based on the observed total score, x = the observed total score, \bar{x} is the average total score for the sample and c is a constant that is based on the correlations of the subscores, as well as the reliabilities and standard deviations of both the subscores and total scores.

To evaluate whether subscores provide added value over the total score, Haberman (2008) suggested evaluating the proportional reduction in mean squared error (PRMSE). The PRMSE is conceptually similar to a reliability coefficient, ranging from 0 to 1; however, as noted by Sinharay (2010), the PRMSE can exceed 1 when the disattenuated correlations among the subscores exceed 1. Hence, a predictor with a larger PRMSE will provide more accurate diagnostic information than a predictor with a smaller PRMSE. The PRMSE (PRMSE_s) for the predictor of the observed subscore, s_s ,

has been shown to be equal to $\rho^2(s_t, s)$, the subscore reliability (for computational details see Haberman, 2008).

The PRMSE (PRMSE_x) for the predictor of the observed total score, s_x , is equal to:

$$\rho^2(s_t, s_x)\rho^2(x_t, x), \quad (26)$$

where $\rho^2(x_t, x)$ is the total test reliability. However, the calculation of $\rho^2(s_t, s_x)$ is computationally more involved,

$$\rho^2(s_t, s_x) = \frac{[Cov(s_t, x_t)]^2}{Var(s_t)Var(x_t)}, \quad (27)$$

where $Cov(s_t, x_t)$ is equal to the sum of the corresponding row taken from the covariance matrix (See Sinharay, Puhon, & Haberman, 2011). The terms in the denominator are as follows:

$$Var(s_t) = \sigma_{s_x}^2 \times \rho^2(s_t, s), \quad (28)$$

where $\sigma_{s_x}^2$ = the observed variance of the subscore, and $\rho^2(s_t, s)$ = observed subscore reliability.

$$Var(x_t) = Var(x) \times \rho^2(x_t, x), \quad (29)$$

where $\rho^2(x_t, x)$ = total score reliability. $Var(x)$ is equal to:

$$Var(x) = \sum_1^n \sigma_{s_x}^2 + \sum_1^p cov(s_x, s_{x'}), \quad (30)$$

where, n = number of subscores, $\sigma_{s_x}^2$ = observed score variance, p = number of subscore pairs, s_x = an observed subscore and $s_{x'}$ = an additional subscore.

Upon calculating the proportion reduction in mean square error for both the subscore and total score predictors, PRMSE_x and PRMSE_s are directly compared. If PRMSE_s is larger than PRMSE_x, there is evidence that the observed subscore is a better predictor of the true subscore than the observed total score. A larger PRMSE_s can also be reconceptualized to

represent a better prediction between the observed subscore and a parallel-form subscore (Sinharay, 2013). In either case, a larger $PRMSE_s$ would suggest that the subdomain score of interest provides accurate diagnostic information about the examinee. However, if $PRMSE_x$ is larger than $PRMSE_s$, then one would conclude that the subscore did not provide added value over the total score as the observed total score provided more accurate diagnostic information (Sinharay, Puhan, & Haberman, 2010).

2.5.1.2 Added Value Based on Classifications

Sinharay (2014) extended Haberman's (2008) method for assessing whether subscores provide added value with respect to examinee classification. Within this approach, it is assumed that the joint distribution of the subscore and the corresponding subscore on a parallel form is approximated by a bivariate normal distribution. For this distribution, the estimated correlation between the forms is equal to $PRMSE_s$ as the correlation between corresponding subscores on two parallel forms is the subscore reliability. The estimated probability that an examinee passes a subtest on both of the parallel forms (P_s) is equal to:

$$P_s = \int_{y=q_s}^{\infty} \left[1 - \Phi \left(\frac{q_s - yr_1}{\sqrt{1-r_1^2}} \right) \right] \phi(y) dy, \quad (31)$$

where

$$q_s = \frac{(c_s - \bar{x}_s)}{s_s}, \quad (32)$$

where c_s is the cut score for classification, \bar{x}_s is the sample mean of the subscore, s_s is the standard deviation of the subscore, and r_1 is equal to $PRMSE_s$ (for further details the reader is referred to Abramowitz & Stegun, 1964). The estimated probability that an examinee fails a subtest on both of the parallel forms (F_s) is equal to:

$$F_s = P_s + 2\Phi(q_s) - 1, \quad (33)$$

which leads to the estimated probability of the same classification across parallel forms being equal to:

$$CC_s = P_s + F_s = 2[P_s + \Phi(q_s)] - 1, \quad (34)$$

where CC_s can be conceptualized as classification accuracy.

To assess whether a subscore has added value with respect to classification, one must compute classification accuracy of the total score. To do this, one must choose an appropriate cut score for the total score, which Sinharay (2014) represented as the same percentile of the sample total-score distribution (c_s). Therefore, the probability that an examinee passes the total test on the original form and the subtest on a parallel form is:

$$P_t = \int_{y=q_s}^{\infty} \left[1 - \Phi\left(\frac{q_t - yr_2}{\sqrt{1-r_2^2}}\right) \right] \phi(y) dy, \quad (35)$$

where r_2 is the estimated correlation between the total score on the original form and the subscore on a parallel form, which is equal to $\sqrt{\text{PRMSE}_s \text{PRMSE}_t}$. The probability of the same classification from the total score on the original form and the subscore on a parallel form is:

$$CC_t = P_t + F_t = 2[P_t + \Phi(q_t)] - 1, \quad (36)$$

where

$$F_t = P_t + 2\Phi(q_t). \quad (37)$$

Therefore, to assess whether a subscore has added value with respect to classification, one can compare CC_s and CC_t . That is, if CC_s is larger than CC_t , one can conclude that a subscore has added value with respect to classification.

Sinharay (2014) applied this procedure to data collected for 4,000 examinees on the TerraNova test, which has five main content areas that include language (34 items), mathematics (57 items), reading (46 items), science (40 items), and social studies (40 items). Twenty cut scores were created across the 1st to 99th percentile of the sample distribution for each subscore and subscore added value was evaluated for each cutscore. Results demonstrated that inferences regarding added value were consistent across all cut scores, except for those at the extremes (e.g., the 1st and 95th percentiles). This result indicates that Haberman's (2008) method would make the same inferences as Sinharay's (2014) method except for at the extreme cut scores. However, it is argued that in practice cut scores at the extreme levels are of little concern, particularly as there is generally less measurement precision at those points of the score distribution, which make it difficult to accurately differentiate examinees with extreme scores. Therefore, Haberman's (2008) method appears to provide robust information concerning added value of reporting subscores even in relation to classifications.

2.5.1.3 Assessing Invariance of Subscore Added Value

As professional standards recommend that scores should not be reported for individuals unless comparability of these scores is established (AERA, APA, & NCME, 2014), Haberman and Sinharay (2013) extended Haberman's (2008) method to assess subscore added value invariance. In addition to professional standards, the motivation for this new method is that previous analyses have assumed the validity of diagnostic information is invariant across all subgroups. However, such an assumption is limited as previous research has demonstrated differential subscore performance by gender and ethnic groups (e.g., Livingston & Rupp, 2004; Stricker, 1993). Therefore, as an extension

of Haberman's (2008) method, Haberman and Sinharay (2013) developed a procedure to determine whether inclusion of subgroup information (i.e., subgroup means and reliabilities) improves subscore estimation when compared to ignoring subgroup information. To ascertain whether the use of subgroup information leads to better estimation of the true subscore, $PRMSE_{sg*}$ is compared to $PRMSE_{sg}$. More specifically, the * subscript denotes that the PRMSE includes subgroup information while the PRMSE values without the * subscript denotes that subgroup information is not included, but the estimates are based solely on data from subgroup g . $PRMSE_{sg*}$ is computed as:

$$PRMSE_{sg*} = \left(1 - \frac{\bar{s}_g - \hat{\rho}_{sg}^2(s - \bar{s}_g)}{\bar{s}_g} \right) - \frac{(\hat{\rho}_{sg}^2 - \hat{\rho}_s^2)^2}{\hat{\rho}_{sg}^2} - \frac{-(1 - \hat{\rho}_s^2)^2}{\sqrt{\bar{s}_g}}, \quad (38)$$

where s is the observed score s , \bar{s}_g is the group g mean for observed score s , $\hat{\rho}_{sg}^2$ is the reliability of subscore s in group g , and $\hat{\rho}_s^2$ is the reliability of subscore s in the entire sample. In contrast, $PRMSE_{sg}$ is equal to the subscore reliability for group g . If $PRMSE_{sg*}$ reduces $PRMSE_{sg}$ from 1.0 by 10%, one can conclude that subgroup information leads to improved true subscore estimation, which would require a follow-up analysis to reveal why there is a lack of subscore added value invariance.

2.5.1.4 Added Value of Reporting Subscores at the Aggregate-Level

As noted by Fu and Qu (2012), little research has been conducted to evaluate the validity of reporting subscores at the aggregate-level. To address this shortcoming in the literature, Haberman, Sinharay, and Puhan (2009) extended Haberman's (2008) method for application in assessing the added value of reporting institutional subscores. Within this method, the validity of reporting institutional subscores is based on whether the average institutional subscore (\bar{s}) is a better predictor of the subscore component for the

institution of the examinee (s_I) than the average total score for the institution (\bar{x}). To assess this, PRMSE values for \bar{s} and \bar{x} must be computed. This is done as follows:

$$PRMSE_{\bar{s}} = \rho^2(s_I, \bar{x}) = \frac{\sigma^2(s_I)}{\sigma^2(\bar{s})}, \quad (39)$$

where

$$\sigma^2(s_I) = K^{-1}(M_{ssI} - M_{sse}), \quad (40)$$

where

$$K = NC/(J - 1), \quad (41)$$

where J is the number of institutions, and

$$C = 1 - \sum_{j=1}^J \left(\frac{n_j}{N}\right)^2, \quad (42)$$

where n_j is the number of examinees in institution j , and N is the total number of examinees, and

$$M_{ssI} = (J - 1)^{-1} \sum_{j=1}^J n_j (\bar{s}_j - \bar{s})^2, \quad (43)$$

where \bar{s} is the mean subscore for all examinees, and

$$M_{sse} = (N - J)^{-1} \sum_{j=1}^J \sum_{i=1}^{n_j} (s_{ij} - \bar{s}_j)^2, \quad (44)$$

where s_{ij} is the subscore for examinee i in institution j . The $PRMSE_{\bar{s}}$ is then compared to the $PRMSE_{\bar{x}}$, which is computed as follows:

$$PRMSE_{\bar{x}} = \rho^2(s_I, x_I) \rho^2(x_I, \bar{x}), \quad (45)$$

where

$$\rho^2(s_I, x_I) = \frac{\hat{c}(s_I, x_I)}{\hat{\sigma}(s_I) \hat{\sigma}(x_I)}, \quad (46)$$

where

$$\hat{c}(s_I, x_I) = M_{sxI} - M_{sxe}, \quad (47)$$

where

$$M_{sxl} = (J - 1)^{-1} \sum_{j=1}^J n_j (\bar{s}_j - \bar{s}) (\bar{x}_j - \bar{x}), \quad (48)$$

where \bar{x}_j is the mean overall score for institution j and \bar{x} is the overall mean score across all examinees,

$$M_{sxe} = (N - J)^{-1} \sum_{j=1}^J \sum_{i=1}^{n_j} (s_{ij} - \bar{s}_j) (x_{ij} - \bar{x}_j), \quad (49)$$

and

$$\rho^2(x_l, \bar{x}) = \frac{\sigma^2(x_l)}{\sigma^2(x_l) + \sigma(x_e)/n}, \quad (50)$$

where

$$\sigma(x_e) = (N - J)^{-1} \sum_{j=1}^J \sum_{i=1}^{n_j} \sqrt{(s_{ij} - \bar{s}_j)^2}. \quad (51)$$

Upon computing $PRMSE_{\bar{s}}$ and $PRMSE_{\bar{x}}$, they are directly compared to assess the validity of reporting institutional subscores. More specifically, if $PRMSE_{\bar{s}} > PRMSE_{\bar{x}}$, the average subscore for the institution is a better predictor of the institutional subscore mean, s_l , than does the average total score for the institution.

2.5.1.5 Brennan's (2012) Utility Index

The procedures developed by Haberman and colleagues are motivated by Kelley's (1947) regressed-score estimates (RSEs). However, as noted by Brennan (2012), the regression of true scores on observed scores leads to some fundamental inconsistencies with certain CTT assumptions. For example, due to regression to the mean, high examinee scores are lowered toward the mean, while low examinee scores are increased toward the mean. Additionally, the CTT assumption that an examinee's true score is equal to the expected value over replications of the measurement procedure is untenable for RSE methods. Furthermore, RSE methods assume a linear regression, which is not assumed in CTT. To overcome some of these limitations, Brennan (2012) introduced the

Utility Index to assess subscore added value, which is based purely on CTT, traditional conceptions of reliability, and is not reliant on RSEs.

Brennan's (2012) method considers three observed-score random variables, which include X (i.e., the subscore of interest), Z (i.e., the total score), and Y (i.e., the non- X component of Z). Based on CTT assumptions, these three variables are decomposed into true-score (T) and error (E) random variables:

$$X = T_x + E_x \quad (52)$$

$$Y = T_y + E_y \quad (53)$$

$$Z = (T_x + E_x) + (T_y + E_y). \quad (54)$$

Under CTT assumptions, the reliability of X is:

$$\rho^2(T_x, X) = \left[\frac{\sigma(T_x, X)}{\sigma(T_x)\sigma(X)} \right]^2 = \frac{\sigma^2(T_x)}{\sigma^2(X)}, \quad (55)$$

where X serves as an estimator of T_x . As with Haberman's (2008) method, the question of added value is based on whether Z is a better estimator of T_x than X . Therefore, replacing Z for X in equation 25, it follows that:

$$\rho^2(T_x, Z) = \left[\frac{\sigma(T_x, Z)}{\sigma(T_x)\sigma(Z)} \right]^2, \quad (56)$$

where $\rho^2(T_x, Z)$ is referred to as the index of utility (U), which is an index that quantifies the utility of using Z as an estimator of T_x that ranges from 0 to 1. However, for simplicity of calculation, the author presents an alternative formula for expressing U without true-score parameters, which he shows to be equal to:

$$U = \frac{[\sigma(X, Z) - \sigma^2(E_x)]^2}{\rho_x^2 \sigma^2(X) \sigma^2(Z)}, \quad (57)$$

where, $\sigma^2(E_x)$ is typically estimated as:

$$\sigma^2(E_x) = \hat{\sigma}^2(X)(1 - \hat{\rho}_x^2). \quad (58)$$

U by itself does not provide information regarding the merits of using Z rather than X . As a result, the author developed the following comparative statistic:

$$\tilde{U} = \frac{U/(1-U)}{\rho_X^2/(1-\rho_X^2)}, \quad (59)$$

where \tilde{U} is the relative utility of using Z instead of X . A nice property of \tilde{U} is that $100|1 - \tilde{U}|%$ is the percentage change in the length of the subscore that is needed to obtain a reliability equal to U . More specifically, if $\tilde{U} \leq 1$, the use of the subscore of interest is supported with respect to reliability. However, if $\tilde{U} > 1$, $100|1 - \tilde{U}|%$ indicates the percentage increase in test length that is required for the subscore to obtain a reliability consistent with the total score. Although Brennan's (2012) method provides some nice features, such as calculation of the U statistic as well as a relative utility index that is akin to Spearman's prophecy formula, the author demonstrated that across SAT verbal subscores both his and Haberman's (2008) method led to the same conclusions regarding subscore added value.

2.5.2 Studies That Have Evaluated Subscore Validity for Operational Data

2.5.2.1 Individual-Level Subscore Validity

A number of operational testing programs have applied Haberman's (2008) method to assess the added value of reporting subdomain performance. As an example, Lyrén (2009) examined the utility of reporting subscores on the *SweSAT*, which is the Swedish version of the SAT. In this study, Lyrén applied Haberman's (2008) method to the following five proposed subtests that were administered to as many as 41,530 examinees (analyses were based on multiple test forms): vocabulary (40 items), Swedish Reading Comprehension (20 items), English Reading Comprehension (20 items), Data Sufficiency (22 items), and Diagrams, Tables, and Maps (20 items). Average subdomain

correlations with the total score were found to range from 0.74 to 0.84, while the average subdomain inter-correlations ranged from 0.40 to 0.66. The relatively moderate subdomain inter-correlations and long subdomain test lengths (minimum of 20 items for each subdomain) led to larger PRMSE_s values for four of the five subtests, which provided validity evidence for subscore reporting on these four subdomains.

Sinharay (2014) applied Haberman's (2008) method to data collected from 4,000 examinees on the TerraNova test, which is composed of five main content areas that include language (34 items), mathematics (57 items), reading (46 items), science (40 items), and social studies (40 items). Internal consistency reliability (coefficient α) across subscores ranged from 0.83 to 0.92, while the inter-subdomain correlations ranged from 0.81 to 0.97. In computing the ratio of PRMSE_s, the author found that all subscores, except for the science subdomain, provided added value. Although both Lyrén (2009) and Sinharay (2014) demonstrated subscore added value in operational contexts, not all analyses of subscore utility support the reporting of diagnostic scores.

As an example, Sinharay, Haberman, and Puhan (2007) examined subscore added value for a basic skills test administered to as many as 3,240 (multiple forms were administered) teachers. The test was composed of six subdomains: 1) reading skills, 2) reading application, 3) mathematics skills, 4) mathematics applications, 5) writing skills, and 6) writing application. The number of items per subdomain was not provided, but the overall internal consistency reliability was 0.94 for all test forms. Although, the inter-subdomain correlations ranged from 0.54 to 0.76, larger PRMSE_x values were obtained for all subdomains across all test forms.

Similar findings were obtained by Puhan, Sinharay, Haberman, and Larkin, (2008) who examined subscore added value for six certification tests that differed in content (mathematics, social studies, and foreign languages), item format (purely multiple-choice, purely constructed-response, and a mixture of multiple-choice and constructed-response items), number of subdomains (three, four, and six), as well as subdomain test-length (2 to 30 items). The authors failed to report either subdomain reliabilities or inter-correlations, but did find that across all six tests, subscore added value was found to be lacking with $PRMSE_S$ values smaller than $PRMSE_X$ values by as much as 0.60 points on a 0 to 1 scale; however, as the test length expanded, the magnitude of the difference in $PRMSE$ values decreased, presumably as the subdomain reliabilities improved.

Haberman (2008) examined subscore added value for both the SAT I math and verbal sections for the 2002 administration. The verbal section was comprised of three subdomains, which had test lengths of 19, 19, and 40, respectively. These subdomains were found to possess correlations with the total score that ranged from 0.87 to 0.96. Similarly, the math section was comprised of three subdomains with 10, 25, and 25 items, respectively. The two subdomains that possessed 25 items had correlations with the total score that were equal to 0.95, while the shorter subdomain had a correlation of 0.82 with the total score. For the verbal subdomains, $PRMSE_S$ ranged from 0.72 to 0.84, while $PRMSE_X$ ranged from 0.87 to 0.89, suggesting that the observed total score was a better predictor of the true subscore than the observed subscore. A similar finding was obtained for the math subdomains as the $PRMSE_X$ values (ranged from 0.89 to 0.92) were much higher than the $PRMSE_S$ values (ranged from 0.64 to 0.87).

Haberman (2008) also examined the utility of reporting subscores for the PRAXIS examination, which is comprised of four subdomains with equal test lengths of 25 items: English language arts, mathematics, citizenship and social science, and science. All subdomains had correlations with the total score that ranged from 0.79 to 0.84. Although, the four subdomains had equal test lengths and similar total score correlations, the English language arts ($PRMSE_S = 0.73$; $PRMSE_X = 0.70$) and mathematics ($PRMSE_S = 0.79$; $PRMSE_X = 0.73$) sections were found to have added value, while the citizenship and social science ($PRMSE_S = 0.68$; $PRMSE_X = 0.77$) as well as the science ($PRMSE_S = 0.69$; $PRMSE_X = 0.80$) sections did not.

In a more thorough evaluation of subscore reporting practices, Sinharay (2010) examined 25 operational tests to assess if reporting subdomain scores supplied added value for examinees over the total test score estimate. Across these 25 operational tests, the number of subscores (ranged from 2 to 7), the average subdomain test length (ranged from 11 to 69), the average internal consistency reliability (ranged from 0.38 to 0.92), and the average subdomain inter-correlations (0.42 to 0.77) differed greatly. In examining PRMSE values, the author found that of the 25 tests, only nine were found to have at least one subscore of added value, while just two tests provided added value for all reported subscores. These latter two tests (SAT I and an ELP assessment) were each comprised of two subdomains with long average subdomain test lengths (43 and 69 items), high average subdomain reliability ($\alpha = 0.90$ and $\alpha = 0.92$), and relatively moderate average subdomain inter-correlations (0.68 and 0.70). Overall, the average number of items for the subtests with added value ranged from 24 – 69, the average

subscore reliability ranged from 0.72 – 0.92, and the average disattenuated subscore inter-correlations ranged from 0.71 – 0.90.

To better generalize the results from analyzing operational tests, Sinharay (2010) conducted a simulation study to understand the conditions in which subscore added value is provided. Within this study, the independent variables examined included: 1) number of subscores (2, 3, or 4), 2) length of the subscores (10, 20, 30, or 50), 3) mean subdomain inter-correlations (0.70, 0.75, 0.80, 0.85, 0.90, or 0.95), and 4) sample size (100, 1,000, or 4,000). Results from this analysis led to numerous conclusions regarding the characteristics that are necessary for obtaining subscore added value. For one, as the subdomain test length increased and the subdomain inter-correlations decreased added value is most often obtained. Specifically, if the average number of items in a subdomain is equal to 10, added value is rare only when the average subdomain inter-correlations are 0.70, whereas inter-correlations that are stronger provide no added value. Regardless of test length, subscores rarely have added value when the inter-correlations are equal to 0.90 or higher. If the average subdomain test length is 20 items or higher, added value is largely dependent on the strength of subdomain inter-correlations. More specifically, for length 20 and correlation ≤ 0.75 , subscores have added value 50% of the time, while the same percentage of added value is obtained for a test length of 50 and correlations ≤ 0.85 . Furthermore, the number of subdomains was found to not impact results, while sample size had minimal impact; however, it should be noted that the author did not examine standard errors or the precision of the point estimate (PRMSE) that is used to judge the utility of reporting subscores. These findings mirror the recommendations provided by Sinharay, Haberman, and Puhan (2007) and Haberman (2008), which suggest that

subscores will provide added value when they are reliable and distinct from one another.

2.5.2.2 Subscore Added Value Invariance

One of the major limitations of previous research evaluating the validity of subscore reporting is that such analyses have assumed that the validity of diagnostic information is invariant across all subgroups. However, such an assumption is limited as previous research has demonstrated differential subscore performance by gender and ethnic groups. For example, Stricker (1993) observed gender differences on subtests of the Law School Admissions Test (LSAT) for the logical reasoning subdomain.

Differential gender performance was also found on constructed response tests across Praxis Principles of Learning, Teaching tests for secondary school teachers, and in subject-knowledge tests of social studies, science, and middle school mathematics (Livingston & Rupp, 2004). These results suggest the need to assess the invariance of subscore added value across subgroups. That is, analyzing added value for subscores across all examinees may bias the inferences made from such analyses as subgroups may possess differential inter-subscore correlations. If this occurs, the validity of diagnostic information at the sub-domain level may differ across subgroups.

Haberman and Sinharay (2013) evaluated a new procedure (described previously) that incorporates subgroup information to improve subscore estimation in operational data. The improvement of incorporating subgroup collateral information was evaluated by comparing PRMSE values to those computed by Haberman's (2008) method for four operational tests by ethnic groups. Data for test 1 came from 4,242 examinees that were assessed on two subdomains that were comprised of 205 multiple-choice items in total. Test 2 (N = 1,932) also possessed two subdomains with each subdomain being comprised

of 100 multiple-choice items. Tests 3 and 4 were teacher-certification tests, which had four (total of 120 items) and three (total of 75 items) subscores that were reported, respectively. For these tests, data were collected from 5,270 and 6,643 examinees, respectively. Subscore invariance for tests 1 and 2 were based on five ethnic groups, while tests 3 and 4 were based on four ethnic groups. Across tests, one ethnic group was made up by combining small ethnic groups (less than 100 examinees) and those who did not provide their ethnicity. Results provided two major findings: 1) subscore information did not improve true subscore estimation, and 2) the ethnic groups evaluated possessed subscore invariance across all subdomains and tests. More specifically, the inferences concerning subscore added value were the same across Haberman and Sinharay's (2013) method that incorporates subgroup information and that of Haberman's (2008), which does not. Furthermore, when applying Haberman's (2008) method at the individual group-level, there was no added value for any subscores across ethnic groups; however, when applying the method across all examinees (i.e., not taking into consideration ethnic groups), tests 1 and 2 were found to have added value for one subdomain.

Sinharay and Haberman (2014) extended the work of Haberman and Sinharay (2013) by evaluating four operational tests for subscore added value invariance. These four tests were comprised of a measure of achievement in several disciplines (Test A), an internal English proficiency assessment (Test B), a teacher certification assessment (Test C), and an assessment for prospective teachers in K – 12 (Test D). Test A was comprised of two test forms (Test A1 and Test A2) with both Test A1 (N = 4,242) and Test A2 (N = 1,932) being comprised of 200 total multiple-choice items and three subdomains. In contrast, Test B (N = 14,000) was comprised of four subdomains with a total of 84

dichotomous, Likert, and constructed-response items. Test C was administered to 2,000 examinees and was comprised of two subscores measured by a total of 40 multiple-choice and three constructed-response items. Lastly, Test D (N = 6,643) had a total of 120 multiple-choice items that were divided into four subdomains. Invariance was assessed by ethnicity, language, gender, and both gender and ethnicity for Tests A, B, C, and D, respectively. For both ethnicity and language invariance analyses, one group was comprised of small minorities (ethnic or language) and examinees that did not specify their group membership (i.e., their ethnicity or language). Subscore added value invariance was assessed differently from Haberman and Sinharay (2013) in that PRMSE values were calculated for both the total and individual group samples. If differences were noted between groups, a subscore was determined to possess a lack of subscore added value invariance. The authors also computed augmented subscores using a procedure akin to Wainer et al. (2001) and evaluated their invariance.

Results demonstrated that for Tests A2, B, and C subscore added value invariance was obtained across ethnic, linguistic, and gender groups, respectively. However, for tests A1 and D a lack of invariance was observed for ethnic groups (i.e., no differences in subscore added value were observed for gender groups on Test D) when using Haberman's (2008) method separately across groups. In more closely examining plausible reasons for a lack of invariance, a number of important trends were noted. For one, large differences in subscore means between groups did not always lead to a lack of invariance. Secondly, the authors found that differential item functioning (DIF) was not related to a lack of invariance; however, as noted by the authors, the amount of DIF present was relatively "small." The factor proposed by the authors to have the largest

impact was differences in inter-subdomain correlations. The reason for this is that as the inter-subdomain correlations decrease, $PRMSE_X$ values also decrease (see Haberman (2008) for computational details), which leads to a lack of invariance. An additional trend was that differential inferences related to subgroup added value disappeared when applying the augmentation procedure that is akin to Wainer et al.'s (2001) procedure. As a result, the authors suggest that in the future testing programs interested in providing diagnostic information should report augmented subscores.

2.6 Summary of Literature Review

A number of subscore estimation methodologies have been proposed to meet both the increased demand for diagnostic information and the psychometric challenge of providing reliable scores that are often based on short subdomain test lengths. These estimation procedures can be categorized into: 1) simple and 2) augmented approaches. Simple approaches are comprised of reporting number or percent-correct scores and estimating abilities using unidimensional IRT estimation (either independently estimating subdomains separately or using fixed item parameter estimation [based on estimates from calibrating all items simultaneously] and estimating ability based on items that belong to the subdomain of interest). However, simple approaches are limited in that they do not address the issue of low subdomain reliability. To address this concern, researchers have developed augmented procedures, which use collateral information from total or other subdomain scores to improve reliability of the subdomain ability estimates. These procedures can be categorized as either CTT (e.g., Kelley's [1947] regressed-score estimate method, and Wainer et al.'s [2001] method) or IRT (e.g., the Objective

Performance Index Yen's [1987], Wainer et al.'s [2001] method, the Out-of-Scale Information method [Kahraman & Kamata, 2004], and MIRT models) procedures.

To assess which augmentation procedure provides the most accurate ability estimation and improved reliability, researchers have conducted multiple comparative studies. Overall, a number of conclusions can be drawn from these studies. For one, when the subdomain test lengths are sufficiently long (e.g. 30 items or more) all subscore procedures perform similarly (Fu & Qu, 2012). Secondly, when subdomain lengths are shorter, augmentation procedures have been shown to improve subscore reliability, particularly when subdomain inter-correlations are high (e.g., $r = 0.90$) when compared to number or percent-correct scores; however, improvements in reliability decrease as the inter-subdomain correlations decrease ($r = 0.30$ to $r = 0.60$; Edwards & Vevea, 2006). Thirdly, although MIRT models are theoretically appealing, they have not demonstrated improved performance in reducing subdomain estimation error when compared to other CTT or UIRT augmentation methods (Luecht, 2003). Lastly, when applying straightforward methods, such as the OPI and Wainer et al.'s (2001) method using both raw and theta scores, reliability is improved by making individual subdomain scores nearly identical for each examinee or making the subscore profiles more similar to the overall sample score profile (Skorupski & Carvajal, 2010). Such a finding suggests that improvements in subdomain reliability may come at the cost of subdomain distinctiveness, which leads to a loss in diagnostic information at the individual-level.

Although augmentation procedures have been shown to improve subscore reliability *at a cost*, analyses of score reporting in practice show that nearly all testing programs reviewed reported subscores as number or percent-correct. Furthermore, when

reporting subscores, testing programs were found to rarely report subscore precision. As a result, it is of little surprise that a number of issues arose when evaluating subscore added value for operational testing programs. For one, the majority of tests reviewed in the literature were found to lack added value across all subscores reported, due to either strong subdomain inter-correlations ($>.90$) or low subdomain reliabilities as subdomain test lengths were short (less than 20 items). Sinharay (2010) conducted simulation analyses to better understand the conditions in which subscore added value is present and found that when the inter-correlations were ≥ 0.90 , added value was rarely present. In contrast, when the subdomain test lengths increased to around 20 items and subdomain inter-correlations ≤ 0.70 , added value was obtained at a rate of 50%.

Previous analyses of subscore added value have been limited in that they have assumed that the validity of diagnostic information is invariant across all subgroups. Such an approach assumes that the subdomain reliabilities and the inter-subdomain correlations are equal across all sub-populations within the examinee pool. To address these possible issues and to ensure score comparability, Haberman and Sinharay (2013) developed an extension of Haberman's (2008) method to incorporate subgroup information to improve subscore estimation. Although the results demonstrated that subscore estimation did not improve with such information, Sinharay and Haberman (2014) did find that across a number of tests, subscore added value invariance was found to be lacking for a number of ethnic and linguistic groups. These findings point towards the need to evaluate subscore added value for examinee subgroups.

The current approach proposed by Sinharay and Haberman (2014) for evaluating the utility of reporting diagnostic information based on manifest characteristics assumes

that measurement models are invariant across individuals within demographic subgroups. However, as Reise and Hidaman (1999) suggest, "...models are good (i.e., fit well) for some people, some of the time, and there simply is no such thing as a ... model that adequately represents important psychological phenomena equally well for all individuals in a given population" (p. 4). Such a statement holds true to subscore reporting as the added value of diagnostic information should not be equal across all individuals within a demographic subgroup, but instead the need for such information should be based on test performance. To this end, it is proposed that multivariate outlier and non-parametric person-fit statistics are applied to individual-level data to identify aberrant score profiles and response patterns respectively due to multidimensionality. This approach may allow for both the detection of examinees that need diagnostic information as well as the ability to provide valid subscores for these individuals.

CHAPTER 3

METHODOLOGY

3.1 Overview

As the goal of this study is to find an alternative approach to reporting distinct raw subscores, the utility of applying a general multivariate outlier detection method as well as a non-parametric person-fit statistic were evaluated for assessing divergence from a unidimensional model at the individual-level. Ideally, such approaches will allow for both the identification and evaluation of subscore added value invariance of unobservable groups (i.e., groups not based on demographic similarities) that differ in the underlying dimensionality of the assessment administered. The effectiveness of these general approaches were investigated in terms of Type I error, power, as well as recovery of added value classifications based on Haberman's (2008) method when manipulating the proportion of examinees with multidimensional score profiles, the degree of multidimensionality (based on a correlated-traits model), and subdomain test lengths. Specifically, the following research questions were addressed:

1. How multidimensional do data need to be for subscores to have added value (i.e., be better predictor of the true subscore than the total score)?
2. As data depart from unidimensionality, how well do the Mahalanobis Distance and H^T person-fit indices identify aberrant score profiles and patterns with respect to Type I error, power, and recovery of descriptive statistics?
3. When separating examinees into groups based on score profiles or response patterns that may differ significantly from the total sample, under what conditions

does subscore added value invariance hold [based on Haberman's (2008) method]?

These research questions were investigated via a number of simulation analyses, while the practical utility of applying procedures to flag aberrant score profiles or response patterns due to multidimensionality were evaluated for a large-scale high-stakes assessment. It should be noted that “added value” is defined in this study using Haberman's (2008) definition which claims subscores have added value when the observed *subscore* of interest is a better predictor of the respective true subscore than the observed *total* score. The sections that follow describe in detail the methodological procedures that were implemented.

3.2 Simulation Study

3.2.1 Data Generation

Data were generated separately for two groups administered an n multiple-choice item test comprised of four subdomains. The two groups simulated in this study differed on the degree of multidimensionality underlying the ability estimates on the four subdomains. Specifically, Group 1 possessed a unidimensional representation of the four subdomains by having inter-subdomain correlations of 1, while Group 2 possessed inter-subdomain correlations that ranged from weak to moderate (using Cohen's, 1968 criteria). To accomplish this, ability estimates (thetas) were sampled from a multivariate standard normal distribution for each simulee as:

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (60)$$

where $\boldsymbol{\theta}$ is a 4 x 1 vector of ability estimates, $\boldsymbol{\mu}$ is a 4 x 1 vector of zeros, and $\boldsymbol{\Sigma}$ is a 4 x 4 covariance matrix with the diagonal components equal to 1 and the off-diagonal

components equal to the inter-subdomain correlations. Item response probabilities were then generated from a four-factor measurement model:

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \quad (61)$$

where \mathbf{x} is an $n \times 4$ matrix of manifest variables, $\mathbf{\Lambda}_x$ is an $n \times 4$ matrix of lambda coefficients, which are the magnitudes of the expected change in the observed variable for a one unit change in the latent variable, $\boldsymbol{\xi}$ is a 4×4 variance-covariance matrix for the latent scores, and $\boldsymbol{\delta}$ is an $n \times 4$ matrix of delta coefficients, which are errors of measurement for the manifest variables and are assumed to be uncorrelated.

As a dichotomous factor analysis model can be viewed as a reparameterization of an IRT model (Kamata & Bauer, 2008), IRT parameters from a three-parameter logistic (3PL) model obtained from the Massachusetts Comprehensive Assessment System (MCAS; Massachusetts Department of Education, 2013) were transformed to obtain the respective slope and intercept parameters:

$$a_i = \frac{\lambda_i}{\sqrt{1-\lambda_i^2}} \quad (62)$$

$$b_i = \frac{\tau_i}{\sqrt{1-\lambda_i^2}} \quad (63)$$

(across models the pseudo-guessing parameter is equivalent). The choice of the generating 3PL model was based on its popularity in the field of educational assessment, while the choice of employing item parameters from an operational testing program was made to ensure that the simulation reflected realistic conditions. Data generation was conducted separately for the two groups of examinees by differing the theta covariance matrix within the *simdata* function in the *R MIRT* package (Chalmers, 2014). To control for sampling error, 25 datasets were generated for each group in every condition. This

resulted in two $N_g \times I$ matrices for each replication, where N_g is equal to the sample size for the respective unidimensional and multidimensional simulee groups and I is the total number of items. Upon obtaining both matrices, they were combined to create one matrix in R (R Core Team, 2014) for identifying outliers that deviate from the unidimensional model and assessing subscore added value.

3.2.2 Degree of Masking Effects when Assessing Subscore Added Value

To assess the degree of masking subscore added value when various proportions of a sample possess multidimensional data, Haberman's (2008) procedure was applied to the total sample for diagnostic purposes. For descriptive purposes, subdomain inter-correlations, total reliability, subdomain internal consistency reliability, as well as $PRMSE_S$ and $PRMSE_X$ values were reported as an average across all subdomains and replications. As mentioned earlier, if $PRMSE_X > PRMSE_S$ when a proportion of the sample possesses multidimensional subdomain scores, there is evidence to both demonstrate masking effects as well as to point towards the need to assess individual model fit.

3.2.3 Assessment of Aberrant Score Profiles and Patterns at the Individual-Level

As previous research has demonstrated that subdomain inter-correlations are often very high when evaluating subscore added value for the total sample (Sinharay, 2010), the objective of this study was to identify unobservable subgroups that differ in the underlying dimensionality of the assessment administered. Such an approach may allow for the distinction of examinees based on whether a unidimensional model best fits the observed item covariances or whether a multidimensional model³ is a better

³ For the purposes of this study, once identifying simulees with poor fit to a unidimensional model, differentiation between multidimensional models was not of concern.

representation. To identify these unobservable groups, indices that assess deviation from an average subdomain vector or score pattern were implemented. Within the literature, person-level model fit indices have been approached from two major frameworks: 1) identification of multivariate outlier cases and 2) person-fit indices (both parametric and non-parametric).

3.2.3.1 Identification of Multivariate Outlier Cases

An outlier or an observation that differs markedly from other observations within a data sample can adversely lead to model misspecification, biased parameter estimation, and incorrect results (Ben-Gal, 2005). Although outliers are most often viewed as error due to clerical mistakes, intentional or motivated mis-reporting, sampling, or faulty distributional assumptions, they may also serve as observations that carry important information or lead to further inquiry (Osborne & Overbay, 2004). One of the most popular methods for identifying observations that are located far from the center of the data distribution (multivariate outliers) is *Mahalanobis Distance*, which is computed as:

$$M_i = \left(\sum_{j=1}^n (x_j - \bar{x}_n)^T V_n^{-1} (x_j - \bar{x}_n) \right)^{1/2}, \quad (65)$$

where n is equal to the number of observations, x_j is a vector of data points for individual j , \bar{x}_n is the sample mean vector, and

$$V_n = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x}_n)(x_j - \bar{x}_n)^T. \quad (66)$$

As an inherent assumption of this test is multivariate normality, cases with large M_i values can be classified as outliers based on some nominal error rate α from a chi-square distribution.

One of the major issues of using Mahalanobis Distance as a criterion for identifying multivariate outliers is that it is susceptible to both masking and swamping

effects. Masking effects occur when the mean and covariance estimates are skewed toward a group of outlying observations, which makes the distance between the outlying observations and the mean small. In contrast, swamping effects occur when non-outlier observations are made to look like outliers by a group of outlying observations skewing the mean and covariance estimates toward the non-outlying observations. As a result, the mean and covariance estimates are clearly biased and can lead to Type I errors and/or low power (Pek & MacCallum, 2011). To overcome this, researchers have proposed a number of robust estimates of multivariate location and scatter (Ben-Gal, 2005). As an example, Hadi (1992) proposed replacing the mean with median and computing the covariance matrix for a subset of observations with the smallest Mahalanobis Distances. Caussinus and Roiz (1990) proposed a robust covariance matrix estimator by weighting observations according to their distance from the centroid. However, of these robust estimators, the most popular used in practice is the minimum covariance determinant (MCD) procedure proposed by Rousseeuw (1984). The objective of the MCD procedure is to compute a mean and covariance matrix based on h ($h < N$) cases that minimize the determinant of the covariance matrix (Hardin & Rocke, 2004). As the MCD method consists of an iterative process, Rousseeuw and Van Driessen (1999) developed an algorithm that improves the speed of computation making it attractive for operational use. Upon computing the mean covariance matrix using the h cases, Mahalanobis Distance can be calculated for each examinee.

For the purpose of this study, cases that depart from unidimensionality were evaluated using the robust Mahalanobis (with MCD estimator) Distance measure. However, instead of evaluating outliers by inputting responses separately for all

dichotomous items, raw subdomain scores were computed and evaluated. This was done for two reasons. First, with non-normal data, such as that with dichotomous independent variables, the Mahalanobis Distance measure can exhibit odd behavior. As an example, greater weight is given to variables with probabilities near zero or one than to variables with probabilities closer to one half (Rosenbaum, 2009). Secondly, if the inter-subdomain correlations were near 1 for the majority of the sample, one would expect to see a relatively flat mean score profile across subdomains. However, for cases that possess a certain degree of multidimensionality, one would expect to see a deviation from the flat score profile, which if great enough, may be detected as an outlier. Computation of the MCD mean and covariance matrix was conducted using the *robust* package in *R* (Wang et al., 2014). Statistical significance of the Mahalanobis Distance measure was based on a chi-square distribution with 4 degrees of freedom at an alpha-level of 0.05 (critical value = 9.488).

3.2.3.2 Person-Fit Indices

The study of examinee-level item score patterns has long been a tradition in evaluating measurement inaccuracy. Such an approach is referred to either as appropriateness measurement or person fit methods, which consist of statistical procedures for assessing the misfit of an individual's test performance to other item-score patterns or an IRT model (Meijer & Sijtsma, 2001). Detection of an atypical person-fit score indicates that the examinee's score pattern cannot be adequately described with the chosen model (Tendeiro & Meijer, 2014). The assessment of person-fit has largely been developed for two types of procedures: 1) non-parametric and 2) parametric (IRT). In general, non-parametric procedures evaluate person-fit based solely on observed scored

responses, while parametric approaches define fit based on the distance between an individual score pattern and the estimated response pattern predicted by an IRT model given estimated parameters.

In a review of person-fit indices, Karabatsos (2003) identified 11 non-parametric and 25 IRT-based person-fit statistics (for a review refer to Meijer & Sijtsma, 2001).

Although numerous person-fit procedures have been developed, the I_z person-fit statistic developed by Drasgow, Levine, and Williams (1985) is one of the most popular methods used in practice; however, research has demonstrated that it performs poorly in detecting a number of different aberrant behaviors (e.g., Karabatsos, 2003; Tendeiro & Meijer, 2014). In general, a number of studies have found non-parametric procedures to outperform their parametric counterparts. As an example, in comparing cheaters, creative responders, guessing, careless-responding, and random-responding, Karabatsos (2003) found that four of the five best performing person-fit indices were non-parametric. Of the non-parametric procedures, the best index across multiple simulation studies has been the H^T index first proposed by Sijtsma in 1986 (Karabatsos, 2003; Sijtsma & Meijer, 1992; Tendeiro & Meijer, 2014).

Sijtsma's (1986) H^T index identifies examinees that do not comply with the Guttman model, which is a simple model to predict item response patterns based on knowledge of an examinee's total score. This is accomplished by ordering dichotomous items in order of increasing difficulty and assuming that all items are cumulative and unidimensional (Kronenfeld, 1972). According to a perfect Guttman scale, an examinee with a score of 10 out of 20 will have correctly answered the first 10 questions and

incorrectly answered the latter 10 questions. Based on this premise, the formula for the H^T index is as follows:

$$H^T = \frac{\sum_{a \neq b} \sigma_{ab}}{\sum_{a \neq b} \sigma_{ab}^{max}}, \quad (67)$$

where $\sum_{a \neq b} \sigma_{ab}$ is the sum of the observed response pattern covariances of all examinees a and b and $\sum_{a \neq b} \sigma_{ab}^{max}$ is the maximum covariance of the observed response pattern for all examinees a and b . The covariance of two examinees' response patterns is given as:

$$\sigma_{ab} = \beta_{ab} - \beta_a \beta_b, \quad (68)$$

where β_{ab} is the proportion of items correctly answered by examinees a and b , and β_a and β_b is the proportion of items correctly answered by examinee a and b , respectively, while the maximum covariance between two response vectors is

$$\sigma_{ab}^{max} = \beta_a(1 - \beta_b). \quad (69)$$

The H^T index ranges from -1 to 1 with a positive value for examinee a indicating an observed response vector that is similar to all other examinees in the dataset, and a negative value indicating that examinee a 's response vector is dissimilar or aberrant.

Computation of the H^T index was conducted using the *PerFit* package in *R* (Tendeiro, 2014). Although Karabatsos (2003) proposed a critical value of .22 for identifying aberrant responses using the H^T index, Linacre (2012) suggested that such a recommendation is not appropriate for all contexts and proposed the need for further simulation studies. To this end, pilot analyses were conducted for three critical values (<0, ≤.025, and ≤.05) and were evaluated in terms of Type I error and power. Results of the pilot analyses demonstrated superior Type I error and power for the <0 critical value

and thus, was the critical value employed for identifying aberrant responses using the H^T index in the full-scale study⁴.

3.2.4 Added Value Evaluation by Outlier Classification

Upon assessing aberrant responses, data were separated into two groups for which data were or were not flagged by the respective procedure under study (either Mahalanobis or H^T). This allowed for the assessment of subscore added value invariance by comparing $PRMSE_X$ and $PRMSE_S$ values across groups (simulees that had adequate fit to the unidimensional model and those that did not) using Haberman's (2008) procedure described in Section 2.5.1.1. $PRMSE_X$ and $PRMSE_S$ values were reported for outlier observations as an average across all subdomains and replications and were reported separately for each method. It should be noted that subscore added value invariance is largely dependent on both the adequacy of the procedures for identifying aberrant responses as well as the independent variables examined as described below.

3.2.5 Independent Variables

In this study, three independent variables were manipulated: 1) proportion of multidimensionality, 2) degree of multidimensionality, and 3) subdomain test length. Across all conditions, the overall number of simulees was held constant at 10,000. The choice of the overall sample size was made to reflect the number of examinees within a grade-level generally observed at the state-level. To assess the impact of sample size on both identification of ill-fitting simulees as well as subscore added value, three different proportions of multidimensionality were simulated from the overall sample size of 10,000: 0%, 10%, 20% and 30%. Simulating 0% multidimensional profiles in the sample

⁴ Pilot study results related to the H^T index are available upon request.

was done for the sole purpose of evaluating Type I error for each method. The remaining proportions were chosen as they each reflect a minority of simulees with multidimensional score profiles, which may be masked in either overall assessments of subscore added value. Therefore, it is of interest to see whether these cases can be identified as ill-fitting and whether their identification will lead to subscore added value when using Haberman's (2008) procedure. If cases are identified correctly, sample size should not have an impact on the stability of added value inferences as a 10% proportion of multidimensionality would equal 1,000 simulees. As noted earlier, Sinharay and Haberman (2014) only found large standard errors with sample sizes less than 150.

The next independent variable examined was degree of multidimensionality (inter-subdomain correlations), which possessed three levels: 1) 0.3, 2) 0.5, and 3) 0.7. More specifically, across all four subdomains the inter-subdomain correlations were held constant to either 0.3, 0.5, or 0.7. The smallest correlation of 0.3 was chosen to represent the most ideal situation of multidimensionality where the subdomains have minimal relationships. Although the inter-subdomain correlation of 0.5 is only slightly stronger, such subdomain correlations were observed in an English language assessment for various native language groups by Sinharay and Haberman (2014). Lastly, a subdomain inter-correlation of 0.7 was examined. Of the three levels, this correlation was the strongest and has been shown by Sinharay (2010) to have added value only under certain conditions. In terms of identification of poor model fit at the individual-level, it is clear that the degree of multidimensionality will play a large role on power. That is, greater departures from unidimensionality were expected to increase the power of identifying

cases that have multidimensional profiles, while correlations that approach 1 may decrease such power.

Although subdomain test lengths were not expected to have as large of an impact on the power of identifying ill-fitting simulees, it was expected to play an important role on the assessment of subscore added value. More specifically, previous research has demonstrated that added value is rare unless a subdomain is comprised of at least 20 items, due largely to the estimation of CTT reliability (Sinharay, 2010). Although previous research has shown that augmentation procedures can improve reliability for subdomains comprised of a small number of items, it does so at the cost of losing subdomain distinctiveness (Skorupski & Carvajal, 2010). As a result, this study examined the following three subdomain test lengths: 10, 25, and 50 items. More specifically, the subdomain test lengths of the four subdomains were held constant across the three levels listed above, which means that the total test lengths were equal to 40, 100, or 200 items, respectively. In an evaluation of operational tests, Sinharay (2010) and Sinharay and Haberman (2014) found that subdomain test lengths can range from approximately 10 to 70 items. As a result, all of the subdomain test lengths chosen for this study are well within what is to be expected in operational tests.

To summarize, the following independent variables and levels were examined:

- Proportion of multidimensionality: 10%, 20%, and 30%
- Degree of multidimensionality (inter-subdomain correlations): 0.3, 0.5, and 0.7
- Subdomain test length: 10, 25, and 50 items

Fully crossed, this produces a 3 x 3 x 3 design for a total of 27 conditions, while an additional three conditions were added to assess Type I error. All conditions were examined across both the Mahalanobis Distance and H^T measures.

3.2.6 Dependent Variables

The evaluation criteria for this simulation study can be broken into two categories: 1) adequacy of procedures for flagging aberrant responses and 2) recovery of subscore added value classifications. More specifically, the adequacy of assessing individual model fit was assessed in terms of Type I error and power. Furthermore, the impact of these flagging procedures on subscore added value invariance were examined via recovery of the group descriptive statistics (reliability, and inter-subdomain correlations) and subscore added value classifications compared to the known values of the generating data.

3.2.6.1 Type I Error

Type I error was defined as the incorrect identification of poor unidimensional model fit for an individual that knowingly possesses adequate model fit. For each of the methods applied, the proportion of simulees across replications incorrectly flagged as possessing poor model fit was reported. It was expected that Type I error rates would not exceed a nominal alpha level of .05.

3.2.6.2 Power

Power was defined as the proportion of true positives at a nominal alpha-level of .05. For each condition, the percentage of simulees correctly identified as possessing ill-fitting data were reported. An arbitrary criterion for adequate power was set at 80%.

3.2.6.3 Recovery of Subscore Added Value Classifications

Although evaluation of Type I error and power provide a basis for judging the adequacy of the flagging procedures, they do not provide a basis for judging the overall impact of misclassification on practical decisions related to inferences concerning subscore added value. As a result, recovery of reliability, inter-subdomain correlations, and PRMSE values were evaluated. Recovery was operationalized in terms of bias, which was simply defined as the difference between the estimated and known parameters for group g . It provides a measure of systematic error in estimation, and was computed as follows:

$$\text{bias} = \frac{\sum_{r=1}^{100}(X_{gr} - E(X_g))}{r}, \quad (71)$$

where X_r is a descriptive statistic for group g on replication r and $E(X_g)$ is the expected descriptive statistic for group g .

3.3 Application of Aberrant Identification Methods to Real Data

As the simulated aspect of this study produced data that were most ideal for evaluating individual-level model fit (e.g., large number of items, multivariate normality, well discriminating items, and only two groups [based on unidimensional and multidimensional correlated-traits models]), it was important to apply the procedures proposed in this study to real data for comparison. For this purpose, data were obtained from 8,803 examinees administered a high-stakes dichotomous item test. This exam was comprised of four subdomains ranging in length from 8 to 16 items (47 total items), and was found to possess adequate total score internal consistency reliability ($\alpha = .91$).

In the simulated data, individuals classified as aberrant were assumed to be multidimensional as the majority of generated data came from a unidimensional model.

To support the same assumption, it was necessary to assess the test dimensionality of the total sample from the applied data. To this end, confirmatory factor analysis was applied to evaluate subdomain distinctiveness. More specifically, two competing models were evaluated: (a) unidimensional and (b) correlated-traits factor structures. The unidimensional model consisted of all items loading onto one latent variable, while the correlated-traits model consisted of a number of latent variables conceptualized as the subdomains. For the latter model, each subdomain was indicated by the items specified in the test blueprint, and all latent variables were correlated. All models were standardized by setting the latent variable residual variances to 1. Model fit was evaluated based on the following fit indices: comparative fit index (CFI), Tucker-Lewis Index (TLI), and root mean square error of approximation (RMSEA). In this analysis, adequate model fit was indicated by CFI and TLI values $>.95$, as well as RMSEA estimates $<.06$ (Hu & Bentler, 1999). Although these fit indices were originally suggested for use with continuous variables, they have also been found to be accurate with categorical variables (Yu and Muthén, 2001).

As the unidimensional model was nested within the correlated-traits model, direct comparisons were made between models to examine which model provided the best fit to the sample data by evaluating Δ CFI. The Δ CFI index was chosen over the traditional chi-square difference test as the latter method has been suggested to be highly sensitive to sample size, while Δ CFI has been demonstrated in simulation studies to provide stable performance with various conditions, such as sample size, amount of invariance, number of factors, and number of items (Meade, Johnson, & Braddy, 2008). Based on simulation analyses, Cheung and Rensvold (2002) recommended that a Δ CFI $\leq .01$ supports the

invariance hypothesis. That is, if the $\Delta CFI \leq .01$, the two competing models would be statistically equivalent. If this were the case, the most parsimonious model was chosen as the best representation of the sample data.

Upon ensuring that the total sample data adhered to a unidimensional model, the robust Mahalanobis Distance measure and the H^T person-fit statistic were applied to the high-stakes testing data. Specifically, raw subdomain scores were computed for each subscore to analyze the Mahalanobis Distance measure, while dichotomous item responses were evaluated to detect atypical score patterns using the H^T person-fit statistic. Each of these individual model fit procedures were conducted as described in sections 3.2.3.1-3.2.3.3, respectively. Descriptive statistics (mean, standard deviation, internal consistency, inter-subdomain correlations, and PRMSE values) were then computed separately for examinees identified as possessing aberrant and non-aberrant response patterns by flagging procedure. This allowed for two separate analyses of score profiles and subscore added value invariance. A score profile analysis allows for the plotting of subdomain scores to evaluate three types of information: level, dispersion, and shape. Level and dispersion of a score profile is the unweighted average and standard deviation of mean subdomain scores, respectively, while the shape of a score profile can be defined as the rank ordering of subdomain means. All three types of information were implemented in this analysis to provide a gauge of the kinds of scores identified as aberrant by the two detection procedures. Such an analysis was important as there was no formal understanding of the characteristics of the “true” outlier profiles. Upon conducting the score profile analysis, subscore added value invariance was evaluated separately for aberrant and non-aberrant examinees by procedure. Taken together, these two analyses

allowed for the assessment of: (a) whether there were groups of examinees with multidimensional data that were masked and (b) whether the flagging procedures functioned similarly to the simulated conditions.

CHAPTER 4

RESULTS

4.1 Overview of Results Section

This chapter is comprised of two sections. The first section presents results from the simulation analyses, while the second section reports results of the application of the aberrant response detection procedures to a real dataset. Section one was broken down into three sub-sections that were based on the research questions outlined in the introduction and methodology chapters. In particular, the first sub-section describes the degree of masking effects on a minority percentage of simulees with multidimensional subdomains when analyzing subscore added value invariance for the total sample. The second sub-section describes Type I error and power rates for two aberrant response detection procedures (H^T and Mahalanobis Distance indices). The last sub-section reports results on the recovery of descriptive statistics and subscore added value classifications for the H^T and Mahalanobis Distance indices. Upon presenting results of the simulation analyses, subscore added value invariance using both the H^T and Mahalanobis Distance indices was evaluated for a large-scale applied dataset. Results from this analysis are provided in the second section of this chapter.

4.2 Degree of Masking Effects when Assessing Subscore Added Value

One of the objectives of this study was to evaluate whether a minority of examinees with multidimensional score profiles can be masked when assessing subscore added value for the total sample via Haberman's (2008) method.

4.2.1 Conditions for Added Value of Multidimensional Generated Data

Before describing the masking effects related to subscore added value analyses using Haberman's (2008) method, it is important to first discuss the rate of added value

that would be expected solely for multidimensional score profiles based on the generating conditions (subdomain test length and inter-subdomain correlations). Such analyses are particularly important for two reasons: a) minimal simulation research has been conducted to provide recommendations on the necessary conditions for obtaining subscore added value, and b) a lack of added value for the generated multidimensional data may serve as a confound when evaluating masking effects.

Results demonstrated that added value was found to be lacking for a number of multidimensional conditions, particularly when the subdomain test length was equal to 10 items. One reason for the lack of subscore added value with this subdomain test length was due to a large underestimation of multidimensional subscore correlations and clearly, a lower reliability due to the small number of subdomain items. Within the conditions with a subdomain test length of 25 items, subscore added value was obtained at a rate of 100% only when the generating subdomain inter-correlations were equal to .30; however, that rate dropped to 61% and 0% with inter-subdomain correlations of .50 and .70, respectively. In general, subscore added value was obtained at much higher rates when the subdomain test length was equal to 50 items. As an example, added value for the generating multidimensional data was obtained 100% of the time across replications when inter-subdomain correlations were equal to .30 and .50; however, added value was obtained for 59% of replications with generating correlations of .70 (Table 1).

4.2.2 Masking Effects for Minority Percentages of Multidimensional Scores

As not all generating conditions were found to provide added value, only those conditions that possessed 100% added value across replications for the generated multidimensional data were evaluated for masking effects. As mentioned, only three

conditions met this criterion: a) subdomain test lengths of 25 items and inter-subdomain correlations of .30, b) subdomain test lengths of 50 items and inter-subdomain correlations of .30, and c) subdomain test lengths of 50 items and inter-subdomain correlations of .50.

Across both subdomain test lengths of 25 and 50 items, there was 0% added value when the proportion of multidimensional score profiles comprised 10% of the total sample, regardless of inter-subdomain correlations. Although the reliability and estimation of inter-subdomain correlations were improved for a subdomain test length of 50 items, the percentage of added value for a proportion of 20% multidimensional score profiles was at most 1% for the condition with subdomain test length of 50 items and inter-subdomain correlations of .30. In fact, the percentage of replications with added value for the total sample was 0% for a subdomain test length of 50 items and inter-subdomain correlations of .50, regardless of the percentage of multidimensional cases in the total sample. Interestingly, the condition with the highest percentage of replications with added value (3%) possessed a subdomain test length of 25 items, 30% multidimensional scores in the total sample, and subdomain inter-correlations of .30; however, by reducing the percentage of multidimensional data in the sample to 20%, the percentage of replications with added value was 0% (Table 2).

The results described in this section demonstrated that a minority of examinees with multidimensional score profiles can be masked when assessing subscore added value for the total sample via Haberman's (2008) method. Specifically, up to 30% of examinees within a sample can possess multidimensional data with inter-subdomain correlations as low as .30 without being identified. This finding indicates that assessing

added value for the total sample may lead to excluding examinees that may possess data that would allow for reporting valid and reliable subscores. To this end, it is necessary to accurately identify such examinees to ultimately assess subscore added value invariance.

4.3 Type I Error and Power by Aberrant Response Identification Procedure

This study evaluated two procedures for assessing aberrant responses from an underlying unidimensional test structure: 1) Mahalanobis Distance and 2) H^T person-fit. The adequacy of each procedure was judged on adequate Type I error and power rates, which were defined as .05 and .80, respectively, across various conditions, such as subdomain test length, percentage of multidimensional scores in the sample data, and inter-subdomain correlations. Furthermore, the practical implications of employing each aberrant response identification procedure on conclusions related to subscore added value were evaluated by assessing bias of descriptive statistics and subscore added value classifications for aberrant responders when compared to generating data. The findings from these analyses are presented below.

4.3.1 Type I Error

Type I error was defined as the incorrect classification of unidimensional scores as aberrant responses solely for data generated via a unidimensional model. Results demonstrated that across aberrant response identification procedures, Type I error rates were found to differ. Specifically, the H^T index was found to be dependent on subdomain test length with Type I error rates decreasing as the number of items within each subdomain increased. As an example, the highest Type I error rate observed for this index was 2% when the subdomain test length was equal to 10 items. Although below the criterion of 5%, Type I error rates decreased to 0.2% and 0.02% as the subdomain test

lengths increased to 25 and 50 items, respectively. In contrast, the Type I error rates for the Mahalanobis Distance index were found to be independent of subdomain test length as Type I error was held constant at 5% when the number of items within each subdomain was equal to 10, 25, and 50 items. Overall, these results suggest that the indices employed did not classify unidimensional cases beyond the a-priori threshold of 5% (Table 3).

4.3.2 Power

Power was defined as the correct classification of scores generated from a multidimensional model as aberrant responses. Results demonstrated that across aberrant response identification procedures, power rates strongly favored the H^T index. As an example, when subdomain test length was equal to 50 items, power rates of 1.0 were obtained regardless of inter-subdomain correlations or proportion of multidimensionality (e.g., Table 5). Similarly, conditions with subdomain test lengths of 25 items all possessed power rates of .99. When subdomain test length was 10 items, power rates of .99 were obtained for all inter-subdomain correlations and proportions of multidimensionality, except for conditions with 30% multidimensionality. Specifically, a power rate of .97 was obtained for an inter-subdomain correlation of .70 (Table 6), while power rates of .98 were observed when inter-subdomain correlations were .30 and .50.

In contrast to the H^T index, power rates were of greater variability for Mahalanobis Distance, which illustrated three general trends. First, power increased as subdomain test length increased. As an example, power rates increased by at least 10% for each respective condition (proportion of multidimensionality and inter-subdomain correlations) when subdomain test length increased by one level. In increasing subdomain

test length by 40 items, power rates increased by as much as 33%. The second trend observed was that power decreased as the proportion of multidimensionality increased. For instance, power decreased when the proportion of multidimensionality went from 10% to 30% by as little as 9% and as much as 15%. This impact was greatest and least variable when the subdomain test lengths were 10 and 50 items, respectively.

As expected, the last trend illustrated was that power decreased as inter-subdomain correlations increased. The most extreme changes in power were observed when increasing inter-subdomain correlations from .50 to .70. As an example, the largest decrease in power (12%) when increasing subdomain correlations by .20 was obtained for a subdomain test length of 50 items and a proportion of multidimensionality of 20%. In the same condition when increasing subdomain correlations from .30 to .70, power decreased by 21%. As a result of the interaction of these three trends, it is no surprise that the lowest power rate (13%) was obtained with a subdomain test length of 10, 30% multidimensionality, and inter-subdomain correlations of .70, while the highest power rate (61%) was obtained with a subdomain test length of 50 items (Table 7), 10% multidimensionality, and .30 inter-subdomain correlations (Table 8).

4.3.3 Recovery of Subscore Added Value Classifications

Although evaluating Type I error and power is important from a methodological standpoint, the practical consequences of employing each procedure must be viewed in the context of making decisions regarding subscore added value. To this end, bias was assessed in terms of descriptive statistics (mean, standard deviation, internal consistency reliability, and inter-subdomain correlations) and subscore added value classifications. As the H^T index was found to possess nearly perfect power and extremely low Type I error

rates, it was expected that minimal bias of added value classifications would be observed. This was the case for nearly all of the conditions, except for the conditions generated with subdomain test lengths of 25 items and inter-subdomain correlations of .50. In particular, under these conditions, an under-classification of subscore added value was observed for the H^T index with under-classification occurring by as much as 18% for the condition with 10% multidimensionality when compared to the generated data (Table 9). Upon closer examination of this condition, the average total score and variability between the generated ($M = 49.67$, $SD = 11.13$) and aberrant cases identified using the H^T index ($M = 49.52$, $SD = 11.37$) were nearly identical ($d = .01$), while subdomain internal consistency reliability was also identical ($\alpha = .68$). The only difference observed was slightly higher average inter-subdomain correlations for the H^T cases ($r = .36$) when compared to the generated data ($r = .34$), which increased the $PRMSE_X$ values for the H^T cases to .65 when compared to .63 for the generated cases. Although slight, this difference appeared to have a large impact on subscore added value classifications as the $PRMSE_S$ (α) values were very similar to the $PRMSE_X$ values. However, besides these conditions, under-classification was minimal and the percentage of added value was nearly identical to the generated data for the H^T cases.

In contrast to the H^T index, the Mahalanobis Distance measure was found to classify multidimensional cases as aberrant across conditions. To examine why this occurred, bias in descriptive statistics (mean, standard deviation, reliability, and correlation values) between the generated data and the cases identified as aberrant using the Mahalanobis Distance measure was evaluated. In doing so, the group of cases identified as aberrant by the Mahalanobis Distance measure were found to consistently

score lower, be more variable, possess higher internal consistency reliability (due to the increased score heterogeneity), and have weaker observed inter-subdomain correlations when compared to the generated data (Table 10). However, bias of the descriptive statistics was impacted by the generating independent variables in a number of ways.

For one, mean score bias was found to be impacted by inter-subdomain correlations. That is, the bias between the mean scores of the aberrant cases identified using Mahalanobis Distance and the generated multidimensional data increased as the inter-subdomain correlations increased. As an example, for a subdomain test length of 10 items and 10% multidimensionality in the total sample, the effect size difference between the identified aberrant cases and the generated data increased from $.15 SD$ for an inter-subdomain correlation of $.30$ ($N = 3179$) to $.84 SD$ for an inter-subdomain correlation of $.70$ ($N = 2393$). In contrast, the bias of score variability (standard deviations) and in turn, internal consistency reliability was found to be independent of the inter-subdomain correlations as negligible differences were observed in a non-consistent pattern across levels. Instead, the biggest impact on bias of score variability and internal consistency was due to subdomain test length as bias decreased when test length increased. As an example, across all conditions with a subdomain test length of 10 items, aberrant cases identified using the Mahalanobis Distance measure possessed higher internal consistency when compared to the generated data by an average of $.24$ points on a scale from 0 to 1. This average bias decreased for the subdomain test lengths of 25 and 50 items to only $.09$ and $.03$, respectively.

Similarly, conditions with subdomain test lengths of 10 possessed on average greater bias in observed inter-subdomain correlations; however, this finding was

confounded by the generating inter-subdomain correlations and proportion of multidimensionality. Specifically, bias in average observed inter-subdomain correlations was found to consistently increase as both the generating inter-subdomain correlations and proportion of multidimensionality increased. As these independent variables interacted with subdomain test length, it is of no surprise that the condition with the largest bias in observed inter-subdomain correlations (.28 on a scale from -1 to 1) was detected for a subdomain test length of 10 items, generated inter-subdomain correlations of .70, and 30% multidimensionality in the total sample. It was of no coincidence that this same condition produced the lowest power rate (13%) for the Mahalanobis Distance measure when compared to all other conditions in the simulation. Similarly, the condition with the highest power rate (61%; subdomain test length of 50 items, generated inter-subdomain correlations of .30, and proportion of multidimensionality of 30%) produced the lowest bias in average observed inter-subdomain correlations (.01; Table 10).

Taken together, these results suggest that the H^T index identified cases generated from a multidimensional model with nearly perfect accuracy. In contrast, the Mahalanobis Distance measure was far less successful; however, the cases identified as aberrant using this method were found to have greater variability between subdomain scores, increased reliability, and lower observed subdomain correlations. To highlight this finding, the reader is referred to Figure 1 where one sees that the multidimensional cases not identified by the Mahalanobis Distance measure as aberrant from the unidimensional model possessed little variability between subdomain scores, which as demonstrated lowered internal consistency reliability and increased inter-subdomain correlations. In contrast, Figure 2 shows the extreme variation of 10 random cases

identified as aberrant using the Mahalanobis Distance measure, which strongly reflects the multidimensionality of the generating data.

One may interpret this variation as being largely due to random error. To evaluate this plausible interpretation, the mean score profile variability of aberrant cases identified by the Mahalanobis Distance measure was compared to the variability of the non-outlier cases in relation to the standard errors of measurement (SEM) for both groups. As can be seen in Figure 3, two general trends were illustrated: a) the mean SEMs for both the outlier and non-outlier cases were nearly identical across subdomain test length conditions, and b) the difference between the mean score profile variability and the SEMs of both groups increased as the subdomain test lengths increased. Clearly, these trends suggested that as the tests became more reliable (i.e., the subdomain test lengths increased), the observed score profile variability was more than would be expected by random error. In particular, the most telling aspect of Figure 3 was that even when the subdomains were at their most unreliable (i.e., 10 items per subdomain), the difference between the SEM and mean score profile variability for the outlier group was still larger than that of the non-outliers. Overall, these findings suggest that the Mahalanobis Distance measure only identified cases with relatively large departures from unidimensionality and score variability that was greater than would be expected based on random error.

When examining bias related to added value classification, the Mahalanobis Distance measure was found to over-classify added value consistently across all conditions. Specifically, added value was obtained for nearly 100% of replications when evaluating cases identified by the Mahalanobis Distance measure even when the

generated multidimensional data showed no added value. Such a finding is supported by briefly considering how decisions regarding added value are made when employing Haberman's (2008) method. That is, if $PRMSE_s$, which is equal to subdomain reliability, is greater than $PRMSE_x$, which is a value based on inter-subdomain correlations as well as subdomain and total score reliability, one concludes that a subscore provides added value beyond that reported by the total score. As noted by Sinharay (2010), a subscore most often provides added value when there is both high internal consistency subdomain reliability and low inter-subdomain correlations. Therefore, it is of no surprise that the Mahalanobis Distance measure obtained cases that possessed subscore added value at rates near 100% as these cases were observed to have high variability between subscores (leading to high subdomain internal consistency reliability and low inter-subdomain correlations). This finding of added value for the Mahalanobis Distance measure was found to be independent of subdomain test length, inter-subdomain correlations, or proportion of multidimensionality in the total sample (Table 10).

4.4 Application of Aberrant Identification Methods to Real Data

Although the Mahalanobis Distance measure showed some promise in simulation analyses, one may question: (a) whether in applied data there are minority groups of examinees with multidimensional data that may be masked by the unidimensionality of the total sample, and (b) whether the Mahalanobis Distance measure functions similarly in practice. To address these concerns, both the H^T index and Mahalanobis Distance measure were applied to high-stakes testing data that were collected from a large sample. Based on the recommendations of Sinharay (2010), this applied dataset would have little probability of providing added value due to the short average subdomain test length.

When analyzing the total sample ($N = 8,803$) of the applied data, adequate model fit was obtained for a unidimensional model ($\chi^2_{1260} = 7579.87, p < .001, CFI = .945, TLI = .989, RMSEA = .024$) as well as a four-factor correlated-traits model ($\chi^2_{1257} = 7292.48, p < .001, CFI = .947, TLI = .989, RMSEA = .023$); however, the correlations of the latent variables from the four-factor model revealed poor discriminant validity between: subdomains 1 and 2 ($\phi_{12} = .955$), subdomains 1 and 3 ($\phi_{13} = .972$), subdomains 1 and 4 ($\phi_{14} = .960$), subdomains 2 and 3 ($\phi_{23} = .949$), subdomains 2 and 4 ($\phi_{24} = .961$), as well as subdomains 3 and 4 ($\phi_{34} = .972$). A direct comparison of the two models, $\Delta CFI = .947 - .945 = .002$, indicated that both recovered the observed covariance matrix with equal accuracy; however, as the inter-factor correlations were found to be extremely high, the unidimensional model was concluded to be the best model. As a result, it was of no surprise that when analyzing the total sample for added value, none of the four subdomains were found to be better predictors of the true subscores than the total sample (Table 11). Therefore, to assess if the aberrant identification procedures could identify distinct groups of examinees with multidimensional scores that would provide added value, the H^T index and Mahalanobis Distance measure were applied to the same data.

4.4.1 Profile Analysis

The agreement in aberrant case classifications by method is provided in Figure 4. Of the 8,803 examinees, the H^T index identified 147 (1.67%) as aberrant, while the Mahalanobis Distance measure identified 579 cases (6.57%). Although the majority of cases were identified as non-aberrant by either procedure, there were only 22 (0.25%) cases that were classified as aberrant by both the H^T and Mahalanobis Distances indices.

The score profiles for the total sample as well as the cases identified by the H^T and Mahalanobis Distance indices are shown in Figure 5.

An examination of the score profiles shown in Figure 5 show that the shape of the total sample and Mahalanobis Distance profiles are nearly identical, but they differ in that both the elevation (grand mean) and scatter (standard deviation) is lower for the Mahalanobis Distance profile. This can clearly be seen as the profiles are almost perfectly parallel with the Mahalanobis profile significantly shifted downwards in Figure 5. In contrast, the profile of the aberrant cases identified by the H^T index is nearly identical to the total sample. In fact, the mean scores between the total sample and H^T index on subdomains 1, 2, and 3 were observed to have overlapping scores when considering standard errors. The one difference observed between these two score profiles was that the cases identified using the H^T index possessed a much higher mean score on subdomain 1, which clearly led to differential profile shapes (Table 12). Overall, this score profile analysis supports the findings from the simulated analyses by demonstrating that the cases identified as aberrant by the H^T and Mahalanobis Distances indices are different in two respects: a) there is little agreement between the two procedures in classifying aberrant cases, and b) not surprisingly, the aberrant cases identified differed in elevation, scatter, and shape.

4.4.2 Subscore Added Value Analysis

As the two procedures were found to identify cases with very different profiles, the next step was to evaluate how these cases were classified in terms of added value. In examining the difference in mean scores between outliers and non-outliers for the H^T index, only one subdomain was found to have non-negligible differences as shown in

Figure 5. Specifically, on subdomain 1, outliers outscored their non-outlier counterparts by an average of .57 standard deviations (Table 13). Furthermore, as would be expected, data for the non-outliers ($n = 8656$) demonstrated very similar inter-subdomain correlations (r ranged from .62 to .71), subdomain reliabilities (α ranged from .65 to .74), total score reliability ($\alpha = .97$), PRMSE_X (ranged from .97 to .98) values, and conclusions regarding subscore added value (none of the four subdomains provided added value) to the total sample. Interestingly, when analyzing data for the outliers identified by the H^T index, higher inter-subdomain correlations (r ranged from .88 to .92) and subdomain reliabilities (α ranged from .88 to .93) were obtained when compared to the non-aberrant cases. Both the increased inter-subdomain correlations and subdomain reliabilities can be explained as artifacts of the increased variability in the subdomain scores of the H^T cases (Table 11). As a result of the high inter-subdomain correlations, subdomain reliability, and total test score reliability (.91), the total score was found to be a better predictor of the true subscores than the observed subscores across all four subdomains.

As no added value was found for the total sample or aberrant responders identified using the H^T index, the Mahalanobis Distance measure was next applied. In evaluating mean performance differences between groups identified by Mahalanobis Distance, aberrant responders were found to significantly score lower across all four subdomains. Specifically, on average, non-outliers outscored their outlier counterparts by .47 to .65 standard deviations (Table 13). Similar to the simulation analyses, inter-subdomain correlations for aberrant responders were found to be significantly lower than their non-aberrant counterparts. Specifically, inter-subdomain correlations ranged from .72 to .77 for non-aberrant responders, while outlier examinees possessed correlations

ranging from .07 to .30. Interestingly, subdomain reliabilities did not significantly increase across all subdomains for examinees classified as aberrant, which was the case in the simulation analyses. As an example, the subdomain reliabilities for subdomains 2 (13 items) and 3 (16 items) were very similar to those of the non-outlier group. The largest difference in subdomain reliability was observed for subdomain 4 (8 items) in which the internal consistency reliability was equal to .79 for outliers and .63 for non-outliers. However, unexpectedly, the subdomain reliability of subdomain 1 (11 items) dropped from .70 for non-outliers to .51 for outliers. One plausible reason for the decrease in reliability may have been due to reduced variability for aberrant responders ($SD = 2.00$) on subdomain 1 when compared to non-aberrant responders ($SD = 2.53$); however, this finding was a bit of an anomaly as the variability in subdomain scores was generally greater for aberrant examinees (Table 11). In addition to lower inter-subdomain correlations, the outlier group was also found to possess lower total test score reliability ($\alpha = .79$) than the non-outlier group ($\alpha = .91$). Taken together, this led to reduced $PRMSE_X$ values for subdomains 1 (.19), 2 (.44), 3 (.55), and 4 (.49). As a result, the outlier examinees identified by Mahalanobis Distance were found to have subscore added value for all subdomains.

However, one question still remains, was the identification of these aberrant cases largely due to random error? To examine this question the average profile variability for the aberrant cases were compared to the standard errors of measurement at each subdomain. As is shown in Figure 6, the score profiles of the aberrant cases were generally more variable than would be expected based on random error, particularly for subdomains 2, 3, and 4. Subdomain 1 was found to possess less variability than the other

subdomains, which led to lower reliability ($\alpha = .51$) and consequentially a higher standard error of measurement. As a result, before reporting all subdomain scores one should consider the relatively strong relationship between subdomain variability and random error for subdomain 1.

Overall, the results from the applied data analysis supported the findings from the simulated data. For one, the real data application demonstrated that in practice there is a distinct group of examinees with multidimensional data that are masked when analyzing subscore added value for the total sample. Such a finding supports the need to identify individuals that deviate from a unidimensional model as their data may allow for reporting useful information that can pinpoint areas of learning needs. Secondly, although both the H^T and Mahalanobis Distance measures identified examinees that differed from the majority of the sample due to aberrant score patterns and profiles, they identified very different types of examinees, which led to divergent conclusions regarding subscore added value. Of the two, the Mahalanobis Distance measure showed the most promise as it identified outliers that provided subscore added value for all subdomains due to low inter-subdomain correlations and increased subdomain reliability (for three of four subdomains). In particular, both simulated and applied data analyses demonstrated that the subdomain relationships were nearly random, which was to be expected as the aberrant cases identified each possessed differential subdomain performance on one or more subdomains when compared to the others. Such increased variability was most pronounced when the subdomain test lengths were short and as mentioned, led to increased subdomain reliability, which was beneficial particularly for the short subdomain test lengths. Taken together, these results suggest that although the

multidimensional group was found to only compose at most 6.57% of the total sample, on a large-scale, such as at the state-level, that could result in hundreds of examinees receiving valid and reliable diagnostic information that could improve instruction and learning in practice when using the Mahalanobis Distance measure.

Table 1

Subscore added value for generating multidimensional data by condition

Test Length	r (outliers)	r	PRMSE _S	PRMSE _X	% Subscore Added Value
10	.30	.13	.45	.49	0%
	.50	.23	.45	.64	0%
	.70	.31	.45	.78	0%
25	.30	.20	.67	.48	100%
	.50	.34	.67	.63	61%
	.70	.47	.67	.77	0%
50	.30	.24	.80	.48	100%
	.50	.40	.80	.63	100%
	.70	.56	.80	.78	59%

Note. These calculations are based on an average of four subdomains and 75 replications

of 10,000 simulees per replication.

Table 2

Subscore added value for total sample by subdomain test length, proportion of outliers, and inter-subdomain correlations

Subdomain Test Length	% Multidim	Inter-Subdomain Correlations								
		$r_{G1} = 1.00$ $r_{G2} = .30$			$r_{G1} = 1.00$ $r_{G2} = .50$			$r_{G1} = 1.00$ $r_{G2} = .70$		
		PRMSE _S	PRMSE _X	% Subscore Added Value	PRMSE _S	PRMSE _X	% Subscore Added Value	PRMSE _S	PRMSE _X	% Subscore Added Value
10	10%	.45	.97	0%	.44	.96	0%	.44	.99	0%
	20%	.44	.92	0%	.44	.94	0%	.44	.96	0%
	30%	.43	.88	0%	.43	.95	0%	.45	.99	1%
25	10%	.66	.95	0%	.67	.97	0%	.67	.98	0%
	20%	.66	.91	0%	.66	.92	1%	.67	.94	1%
	30%	.66	.85	3%	.68	.91	1%	.66	.95	0%
50	10%	.80	.95	0%	.80	.96	0%	.80	.98	0%
	20%	.80	.89	1%	.79	.93	0%	.80	.96	0%
	30%	.80	.87	1%	.80	.90	0%	.80	.93	2%

Note. The PRMSE values reported are an average of the four subdomains and the highlighted cells denote the conditions that were evaluated as the generated multidimensional data possessed 100% added value across replications. % Multidim = the proportion of multidimensional cases in the total sample.

Table 3
 Type I error rate by aberrant detection procedure

Subdomain Test Length	Aberrant Detection Procedure	
	H^T	Mahalanobis Distance
10	.02 (.98)	.05 (.95)
25	<.01 (>.99)	.05 (.95)
50	<.01 (>.99)	.05 (.95)

Note. The numbers in parentheses denote the percentage of unidimensional cases correctly identified as non-aberrant.

Table 4

Power rate by aberrant detection procedure

Subdomain Test Length	% Multidimensionality	Aberrant Detection Procedure					
		Robust Mahalanobis Distance Measure			H^T Person-Fit Statistic		
		Inter-Subdomain Correlations					
		$r_{G1} = 1.00$ $r_{G2} = .30$	$r_{G1} = 1.00$ $r_{G2} = .50$	$r_{G1} = 1.00$ $r_{G2} = .70$	$r_{G1} = 1.00$ $r_{G2} = .30$	$r_{G1} = 1.00$ $r_{G2} = .50$	$r_{G1} = 1.00$ $r_{G2} = .70$
10	10%	.32 (.68)	.26 (.74)	.24 (.76)	.99 (.01)	.99 (0)	.99 (0)
	20%	.23 (.77)	.21 (.79)	.22 (.78)	.99 (.01)	.99 (0)	.99 (0)
	30%	.17 (.83)	.14 (.86)	.13 (.87)	.98 (.02)	.98 (0)	.97 (0)
25	10%	.43 (.57)	.39 (.71)	.34 (.66)	.99 (.01)	.99 (0)	.99 (0)
	20%	.37 (.63)	.33 (.67)	.26 (.74)	.99 (.01)	.99 (0)	.99 (0)
	30%	.34 (.66)	.28 (.72)	.19 (.81)	.99 (.01)	.99 (0)	.99 (0)
50	10%	.61 (.39)	.51 (.49)	.44 (.56)	1.00 (0)	1.00 (0)	1.00 (0)
	20%	.55 (.45)	.47 (.53)	.35 (.65)	1.00 (0)	1.00 (0)	1.00 (0)
	30%	.50 (.50)	.42 (.58)	.33 (.67)	1.00 (0)	1.00 (0)	1.00 (0)

Note. The numbers in parentheses denote the percentage of multidimensional cases incorrectly identified as non-aberrant (Type II errors).

Table 5

Confusion matrix for the H^T condition with the highest power rate

True Classification	Mahalanobis Distance Classification	
	Unidimensional	Multidimensional
Unidimensional	999 (99.999%)	1 (.001%)
Multidimensional	0 (0%)	9000 (90%)

Note. This condition is based on a subdomain test length of 50 items, an inter-subdomain correlation of .30, and 10% multidimensionality. Numbers in parentheses denote the percentage of total observations identified.

Table 6

Confusion matrix for the H^T condition with the lowest power rate

True Classification	Mahalanobis Distance Classification	
	Unidimensional	Multidimensional
Unidimensional	2937 (29.37%)	63 (.63%)
Multidimensional	206 (2.06%)	6794 (67.94%)

Note. This condition is based on a subdomain test length of 10 items, an inter-subdomain correlation of .70, and 30% multidimensionality. Numbers in parentheses denote the percentage of total observations identified.

Table 7

Confusion matrix for the Mahalanobis Distance condition with the lowest power rate

True Classification	Mahalanobis Distance Classification	
	Unidimensional	Multidimensional
Unidimensional	6,659 (66.59%)	341 (3.41%)
Multidimensional	382 (3.82%)	2,618 (26.18%)

Note. This condition is based on a subdomain test length of 10 items, an inter-subdomain correlation of .70, and 30% multidimensionality. Numbers in parentheses denote the percentage of total observations identified.

Table 8

Confusion matrix for the Mahalanobis Distance condition with the highest power rate

True Classification	Mahalanobis Distance Classification	
	Unidimensional	Multidimensional
Unidimensional	8,554 (85.54%)	446 (4.46%)
Multidimensional	383 (3.83%)	617 (6.17%)

Note. This condition is based on a subdomain test length of 50 items, an inter-subdomain correlation of .30, and 10% multidimensionality. Numbers in parentheses denote the percentage of total observations identified.

Table 9

Bias of PRMSE values and percentage of subscore added value for outlier groups identified by individual indices

Test Length	Condition		Robust Mahalanobis			H^T Person-Fit		
	r (outliers)	% Multidimensionality	PRMSE _S	PRMSE _X	% Subscore Added Value	PRMSE _S	PRMSE _X	% Subscore Added Value
10	.30	10	.20	-.19	99%	.04	.16	1%
		20	.25	-.24	100%	.03	.08	0%
		30	.26	-.28	99%	.01	.06	0%
	.50	10	.23	-.32	100%	.04	.11	0%
		20	.23	-.34	100%	.02	.06	0%
		30	.27	-.38	100%	.01	.04	0%
	.70	10	.23	.45	100%	.04	.06	0%
		20	.23	-.43	96%	.02	.03	0%
		30	.26	.46	96%	.01	.03	0%
25	.30	10	.07	-.08	0%	0	.02	0%
		20	.09	-.14	0%	0	.01	0%
		30	.10	-.17	0%	0	.01	0%
	.50	10	.07	-.17	45%	0	.02	-18%
		20	.08	-.21	47%	0	.01	-5%
		30	.09	-.23	26%	.01	0	-7%
	.70	10	.07	-.26	100%	0	0	0%
		20	.09	-.28	99%	0	0	0%
		30	.10	-.31	100%	.01	.01	0%
50	.30	10	.02	.01	0%	0	.01	0%
		20	.03	-.07	0%	0	0	0%
		30	.04	-.12	0%	0	0	0%
	.50	10	.03	-.07	0%	0	0	0%
		20	.03	-.12	0%	0	0	0%
		30	.04	-.16	0%	0	0	0%
	.70	10	.02	-.13	46%	0	0	0%
		20	.03	-.17	36%	0	0	-1%
		30	.03	-.17	41%	0	.01	0%

Table 10

Bias of descriptive statistics for aberrant responders identified using Mahalanobis Distance

Test	Condition		Generating Data				Mahalanobis Distance				Bias			
	<i>r</i>	%	M	SD	α	<i>r</i>	M	SD	α	<i>r</i>	M	SD	α	<i>r</i>
10	.30	10	4.86	1.92	.46	.14	4.70	2.48	.66	.02	.16	-	-	.12
		20	4.96	1.90	.45	.13	4.76	2.51	.69	.02	.20	-	-	.11
		30	5.03	1.91	.45	.14	4.83	2.56	.71	-	.20	-	-	.20
	.50	10	5.07	1.91	.45	.23	4.74	2.53	.68	.03	.33	-	-	.20
		20	4.96	1.90	.45	.22	4.69	2.51	.68	.01	.27	-	-	.21
		30	4.93	1.90	.45	.23	4.71	2.58	.72	-	.22	-	-	.24
	.70	10	4.99	1.90	.45	.32	4.70	2.54	.68	.06	.29	-	-	.26
		20	4.92	1.90	.45	.31	4.58	2.48	.68	.08	.34	-	-	.23
		30	4.94	1.91	.45	.32	4.64	2.58	.71	.04	.30	-	-	.28
25	.30	10	12.51	3.91	.67	.20	12.32	4.72	.75	.15	.19	-	-	.05
		20	12.52	3.88	.67	.20	12.27	4.65	.76	.09	.25	-	-	.11
		30	12.50	3.89	.67	.20	12.22	4.64	.77	.07	.28	-	-	.13
	.50	10	12.42	3.92	.68	.34	12.18	4.71	.75	.21	.24	-	-	.13
		20	12.44	3.90	.67	.34	12.10	4.63	.75	.17	.34	-	-	.17
		30	12.32	3.93	.67	.34	11.95	4.66	.76	.14	.37	-	-	.20
	.70	10	12.55	3.92	.67	.48	12.15	4.73	.75	.27	.40	-	-	.21
		20	12.51	3.89	.67	.47	12.03	4.71	.76	.25	.48	-	-	.22
		30	12.65	3.91	.67	.47	12.14	4.76	.77	.22	.51	-	-	.25
50	.30	10	25.11	7.16	.80	.24	24.95	8.02	.83	.25	.16	-	-	-
		20	25.04	7.13	.80	.24	24.69	7.96	.83	.17	.35	-	-	.07
		30	24.97	7.11	.80	.24	24.60	8.02	.84	.12	.37	-	-	.12
	.50	10	25.01	7.14	.80	.40	24.76	8.03	.83	.35	.25	-	-	.05
		20	25.30	7.09	.80	.40	24.79	7.91	.83	.28	.51	-	-	.12
		30	24.90	7.11	.80	.40	24.57	7.92	.84	.25	.33	-	-	.15
	.70	10	24.93	7.13	.80	.56	24.41	8.00	.82	.44	.52	-	-	.12
		20	24.94	7.13	.80	.56	24.18	8.00	.83	.40	.76	-	-	.16
		30	24.94	7.13	.80	.57	24.11	7.89	.83	.38	.83	-	-	.19

Note. All values were reported as an average of four subdomains.

Table 11

Analysis of added value by identification of aberrant responses for applied data

Method	Data Type	n	Subdomain	# Items	M	SD	Inter-Subdomain Correlations				PRMSE _S	PRMSE _X	Added Value?
							1	2	3	4			
---	Total	8803	1	11	5.06	2.52	1	.69	.71	.63	.70	.96	No
			2	13	8.07	3.05	---	1	.73	.68	.75	.97	No
			3	16	9.34	3.60	---	---	1	.70	.77	.98	No
			4	8	5.16	1.95	---	---	---	1	.66	.97	No
H^T	Non-Outlier	8656	1	11	5.00	2.45	1	.68	.69	.62	.68	.97	No
			2	13	8.05	3.01	---	1	.71	.66	.74	.97	No
			3	16	9.31	3.53	---	---	1	.69	.76	.98	No
			4	8	5.16	1.92	---	---	---	1	.65	.97	No
H^T	Outlier	147	1	11	6.79	3.70	1	.92	.91	.88	.89	1.00	No
			2	13	7.69	4.37	---	1	.92	.90	.91	1.00	No
			3	16	9.18	5.44	---	---	1	.90	.93	1.00	No
			4	8	4.56	2.94	---	---	---	1	.88	1.00	No
Robust	Non-Outlier	8224	1	11	5.14	2.53	1	.73	.74	.67	.70	1.00	No
			2	13	8.20	2.99	---	1	.77	.72	.74	1.00	No
			3	16	9.52	3.54	---	---	1	.73	.76	1.00	No
			4	8	5.26	1.87	---	---	---	1	.63	1.00	No
Robust	Outlier	579	1	11	3.99	2.00	1	.07	.09	.14	.51	.19	Yes
			2	13	6.40	3.35	---	1	.12	.17	.79	.44	Yes
			3	16	7.02	3.56	---	---	1	.30	.75	.55	Yes
			4	8	3.89	2.44	---	---	---	1	.79	.49	Yes

Table 12

Profile analysis of applied data for total sample and outliers by index

Method	Subdomain	Mean Subdomain Score	Elevation	Scatter	Shape
Total Sample N = 8803	1	5.06			-1.85
	2	8.07			1.16
	3	9.34	6.91	2.14	2.43
	4	5.16			-1.75
H^T Outliers N = 147	1	6.79			-0.27
	2	7.69			0.64
	3	9.18	7.06	1.93	2.13
	4	4.56			-2.50
MD Outliers N = 579	1	3.99			-1.34
	2	6.4			1.08
	3	7.02	5.33	1.62	1.70
	4	3.89			-1.44
H^T & MD Outliers N = 22	1	5.18			1.37
	2	4.13	3.82	1.60	0.32
	3	4.45			0.64
	4	1.5			-2.32

Note. Non-outliers were not listed as the results were nearly identical to the total sample.

Table 13

Mean subdomain differences between aberrant and non-aberrant cases by detection method

Subdomain	H^T Index				Effect Size <i>d</i>
	Aberrant (N = 8656)		Non-Aberrant (N = 147)		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
1	5.00	2.45	6.79	3.70	.57*
2	8.05	3.01	7.69	4.37	-.10
3	9.31	3.53	9.18	5.44	-.03
4	5.16	1.92	4.56	2.94	-.24*

Subdomain	Mahalanobis Distance Index				<i>d</i>
	Aberrant (N = 8224)		Non-Aberrant (N = 579)		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
1	5.06	2.52	3.99	2.00	-.47*
2	8.07	3.05	6.40	3.35	-.52*
3	9.34	3.60	7.02	3.56	-.65*
4	5.16	1.95	3.89	2.44	-.58*

Note. * denotes that the mean score difference between the total sample and respective aberrant identification procedure was statistically significant based on an independent-group *t*-test with an alpha-level of .05.

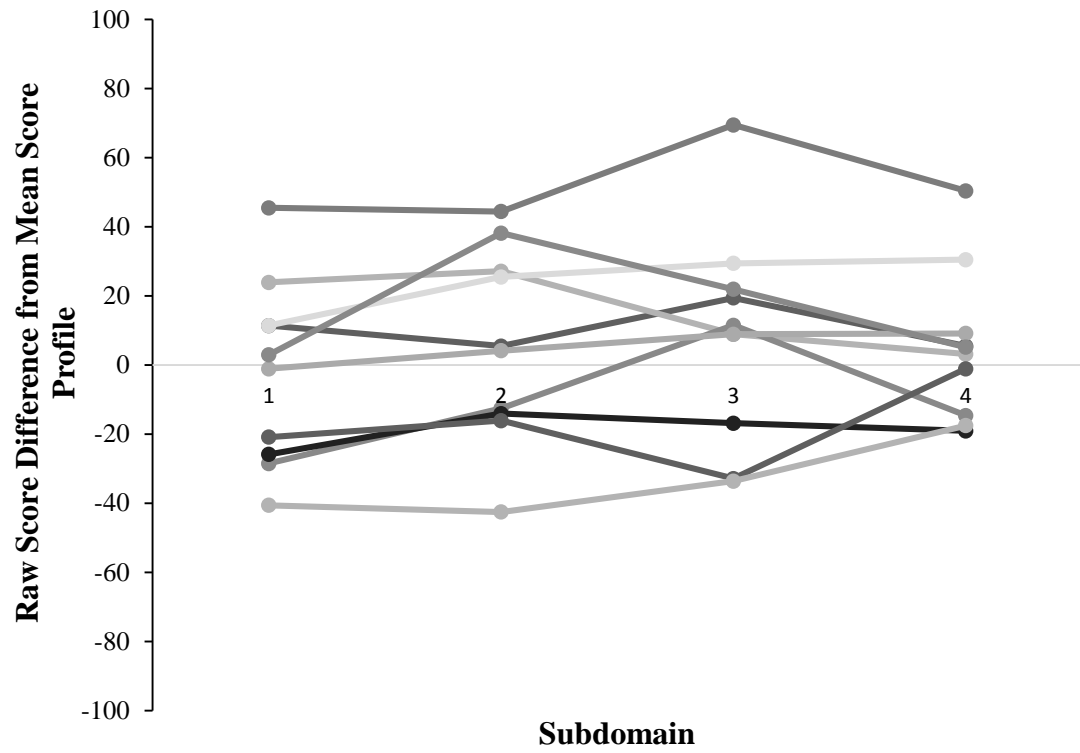


Figure 1. Multidimensional cases incorrectly identified as non-aberrant from the unidimensional model by the Mahalanobis Distance measure.

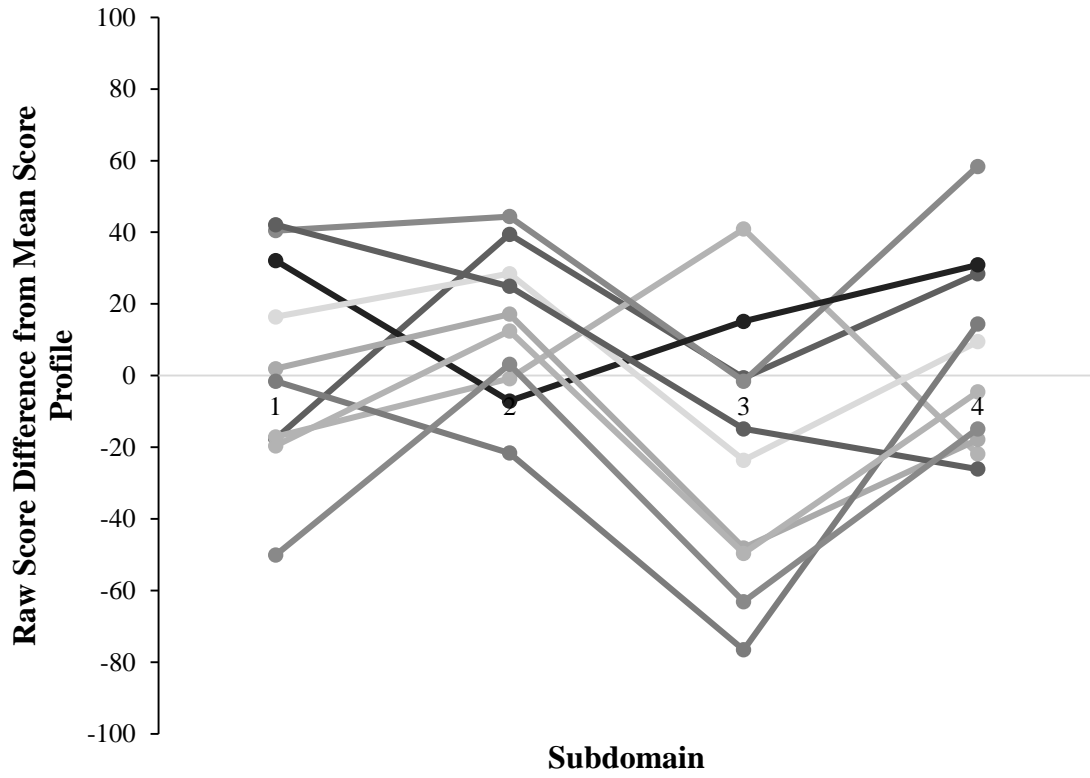


Figure 2. Multidimensional cases correctly identified as non-aberrant from the unidimensional model by the Mahalanobis Distance measure.

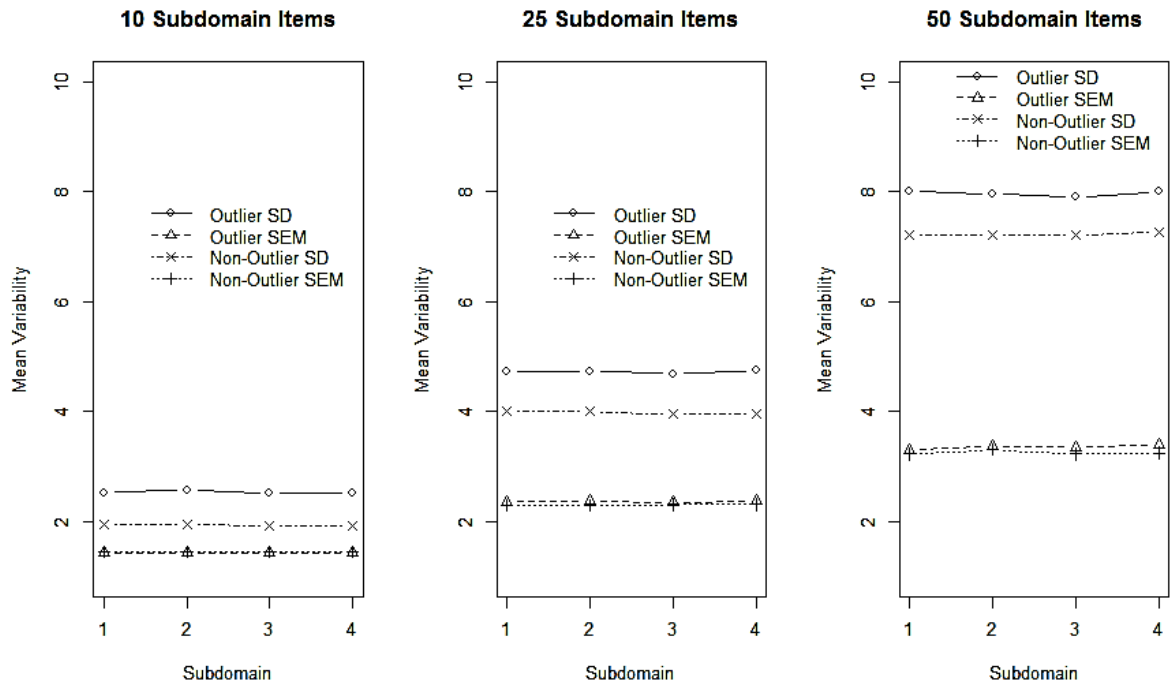


Figure 3. Comparison of mean score profile variability and standard error of measurement for outlier observations identified via the Mahalanobis Distance measure across conditions. Note that the inter-subdomain correlations were equal to .70 and the proportion of multidimensionality was equal to 10% for all three conditions illustrated.

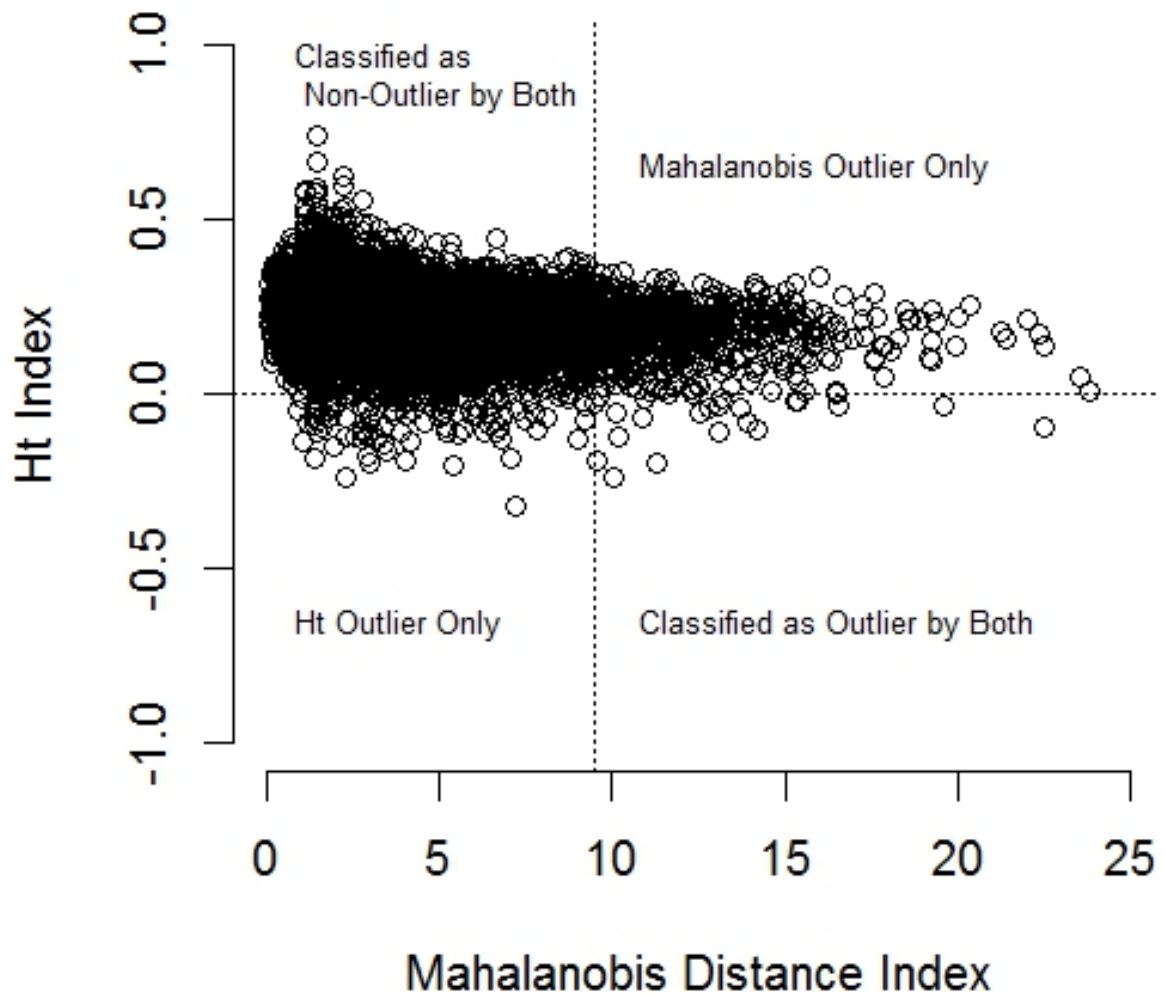


Figure 4. Scatterplot of outlier classifications by the Mahalanobis Distance and H^T indices for the applied dataset.

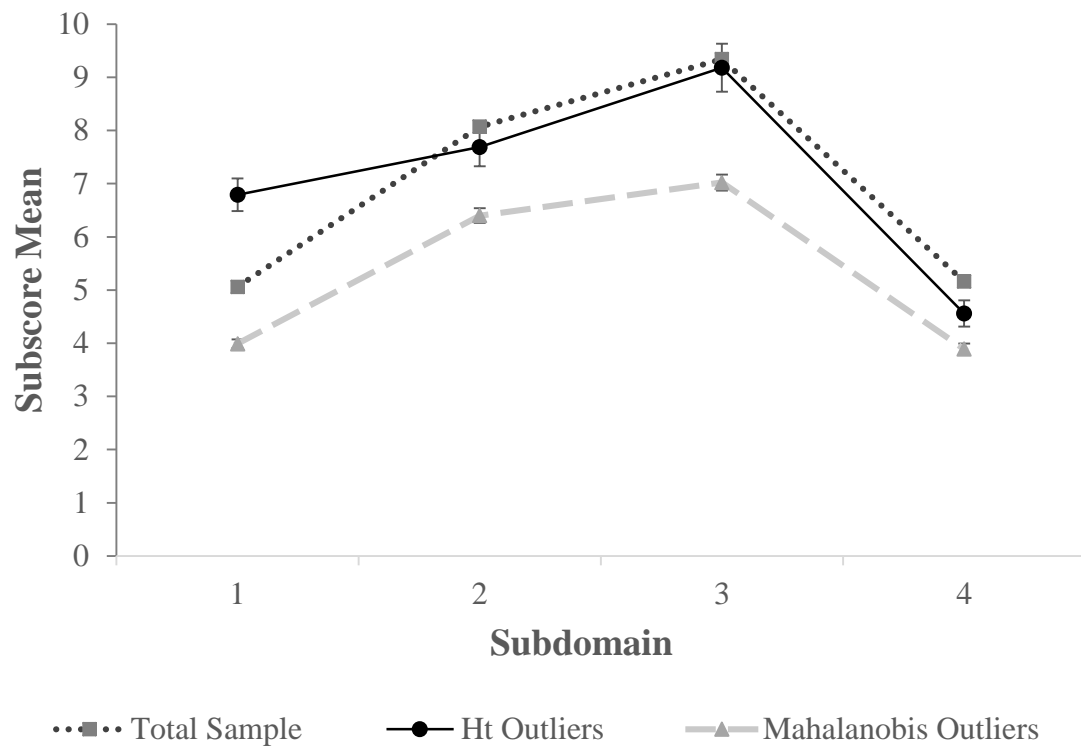


Figure 5. Score profiles of applied dataset for total sample and outliers identified by procedure.

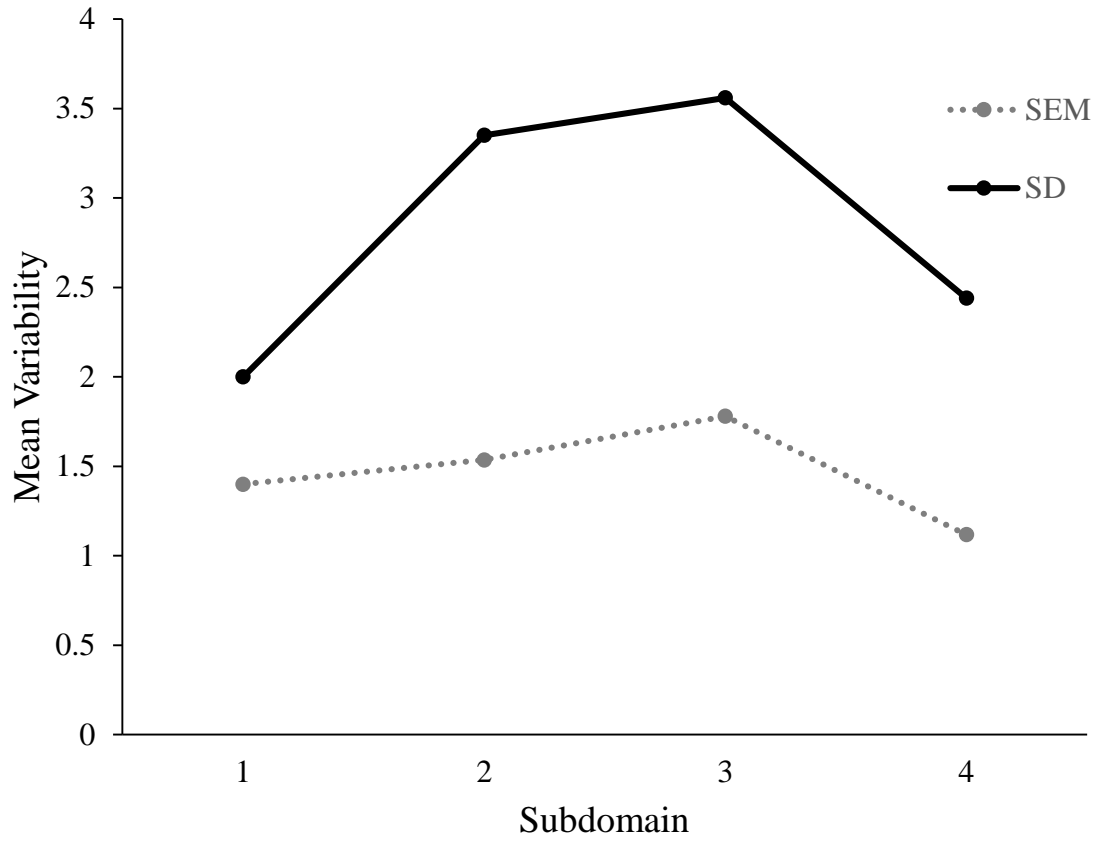


Figure 6. Applied comparison of mean score profile variability and standard error of measurement for outlier observations identified via the Mahalanobis Distance measure.

CHAPTER 5

DISCUSSION

5.1 Overview

The need for educational assessments to serve as a basis for evidence-based decision making for instructional and institutional reform has led to an increased desire for fine-grained feedback. The methodological discussion regarding diagnostic information in the literature has been consumed by developing new methodologies to provide such information in a valid and reliable way. Although there are multiple approaches to obtaining diagnostic information, research into practice has taught us two things: (a) fine-grained level information is most often provided using raw subscores, and (b) these raw subscores rarely provide added value beyond the total score. Such trends suggest that the current practice of reporting raw subscores as diagnostic information may lead to unintended consequences of score interpretation, such as the misplacement of resources in modifying policy and instruction and/or incorrect high-stakes accountability outcomes.

To provide a solution to the current state of diagnostic score reporting, this study proposed a new approach to deciding on who should receive such information. Presently, researchers and practitioners have emphasized that *every* examinee will receive subscores; however, it is argued that diagnostic feedback should be required *only* for examinees that demonstrate a need, which can be defined as an individual that demonstrates poor test performance on one or more subdomains when compared to their performance on the remaining subdomains. Therefore, this study was implemented to illustrate that raw subscores can be valid and reliable for *some* examinees, even when

there is not subscore added value for the total sample due to short subdomain test lengths (e.g., 10 items) and strong inter-subdomain correlations (e.g., $r = .70$). This chapter discusses the major findings of the study and reference is made to the literature where possible. The first section discusses results related to the degree of masking effects on subscore added value when only a proportion of examinees possess multidimensional data. This is followed by a discussion of accuracy rates for two procedures employed to identify cases at the examinee-level that significantly diverged from unidimensionality. The chapter concludes with outlining some limitations of the study and directions for future research.

5.2 Degree of Masking Effects on Subscore Added Value

Evaluations of subscore added value for operational testing programs have rarely demonstrated valid subscore reporting. This has led researchers to largely disparage the use of reporting raw subscores as diagnostic information due to concerns of consequential validity. Regardless of the psychometric concerns, practitioners have continued to demand that subscores are reported, which has caused a rift between the ethical responsibility of measurement professionals and serving the needs of clients or stakeholders (Brennan, 2012). To address the issue of reporting raw subscores as diagnostic information, researchers have proposed the use of augmentation procedures to improve subscore reliability by utilizing collateral information (i.e., the total score or scores from other subdomains) to improve the stability of subscore estimation (e.g., Haberman, 2008; Wainer et al., 2001). However, as noted by Skorupski and Carvajal (2001), these methods largely improve subscore reliability at the cost of subscore distinctiveness by forcing an examinee's score profile to look more like the mean score

profile for all examinees or by making all subscore means and standard deviations for an examinee essentially the same across all subdomains. As augmentation procedures have proved to be less useful than originally thought, a new approach was needed to identify individuals with added value due to subscore distinctiveness.

The new approach conceptualized in this study was derived from the belief that subscore added value is not invariant across all examinees as a model cannot adequately represent mental phenomena equally well for all individuals (Reise & Hidaman, 1999). To support this assertion it was necessary to demonstrate that individuals that differ from the majority of the sample in terms of subdomain distinctiveness can often go unnoticed when assessing added value for the total sample. To this end, simulation studies were implemented to evaluate the degree of masking effects for examinees with multidimensional data when *only* examining subscore added value for the total sample, which is the current practice. Specifically, up to 30% of the total sample was simulated to possess multidimensional data, based on a correlated-traits model, with varying degrees of subdomain inter-correlations (.30, .50, and .70) and subdomain test lengths (10, 25, and 50), while the remaining sample possessed unidimensional data. However, to avoid confounding effects when assessing the degree of masking effects, it was first important to assess the conditions necessary for added value when the total sample possessed multidimensional data.

Although Sinharay (2010) also conducted an analysis to evaluate the conditions necessary to provide subscore added value, this study differed in two ways: (a) much lower inter-subdomain correlations were examined and (b) disattenuated correlations were not used in evaluating added value via Haberman's (2008) procedure. Specifically,

the lowest level of inter-subdomain correlations evaluated by Sinharay was .70, whereas this study looked at inter-subdomain correlations as low as .30 and .50. In terms of disattenuated correlations, they were not used in this study as they can lead to inappropriate inflation of estimates if there is substantial: (a) underestimation of reliability, (b) sampling error, and (c) outliers (Osborne, 2003). As it was expected that in the context of assessing subscore added value that there would be substantial underestimation of reliability due to short subdomain test lengths, use of disattenuated correlations may have confounded the findings of this study and as a result, were not employed.

In terms of examining the conditions necessary for subscore added value, results demonstrated that regardless of inter-subdomain correlations no subscore added value was obtained when subdomain test lengths were equal to 10 (40 total items). As the subdomain test length increased, the percentage of replications with subscore added value increased for most inter-subdomain correlation conditions. As an example, when the subdomain test lengths were equal to 25 or 50 items and the inter-subdomain correlations were equal to .30, added value was obtained for 100% of replications. Similarly, 100% added value was observed for an inter-subdomain correlation of .50 with a subdomain test length of 50, but added value decreased to 61% when the subdomain test length decreased to 25 items. Interestingly, 0% added value was obtained for inter-subdomain correlations of .70 with subdomain test lengths of 25 items; however, when the subdomain test length increased to 50 items, added value was observed for 59% of the replications.

Results from this study differed from Sinharay (2010) in a number of ways. As an example, Sinharay (2010) found that subscore added value was obtained 25% of the time when there were four subdomains, subdomain test lengths were equal to 10, and inter-subdomain correlations of .70 were held constant across subdomains, while this study found added value for 0% of replications. Furthermore, for the condition with four subdomains, inter-subdomain correlations of .70, and subdomain test lengths of 50, Sinharay found added value at a 100% rate, whereas this study found added value at 61%. The higher levels of added value observed by Sinharay may have been due largely to the use of both disattenuated correlations and different generating item parameters. As an example, the average subscore reliability obtained in this study was .45 for a subdomain test length of 10 items and inter-subdomain correlations of .70, while under the same conditions, Sinharay obtained an average subscore reliability of .56. Furthermore, across all conditions, Sinharay only simulated sample sizes of 1,000 for each replication, while in this study 10,000 simulees were included in each replication, which may have been one of the plausible differences in the percentage of added value observed between the two studies. That is, the smaller sample size employed in Sinharay's study may have led to less stability in $PRMSE_S$ and $PRMSE_X$ estimates, which may have led to increased rates of added value. As an example, for the condition with 10 subdomain items and inter-correlations of .70, the average $PRMSE_S$ (.57) and $PRMSE_X$ (.62) values were extremely similar, which could have meant that due to sampling error a number of replications were classified as providing added value when in actuality no added value was provided. Regardless, both studies concluded that added value is rarely provided when subdomain test lengths are as low as 10 items.

This study also contributed to the literature by evaluating added value for inter-subdomain correlations that were much lower than those in Sinharay's study. Specifically, this study showed that under the most ideal situation of multidimensionality where inter-subdomain correlations of .30 were present, no added value was observed when subdomain test lengths were 10 items. Even under more realistic correlations of .50, which have been observed in operational tests by Sinharay and Haberman (2010), added value was only partially observed for subdomain test lengths of 25 items. Taken together, this study provides further evidence that when analyzing subscore added value for the total sample, multidimensional data does not guarantee subscore added value.

The next step was to evaluate whether those examinees that met the necessary conditions for added value could be masked if the majority of the sample possessed unidimensional data. To avoid confounding effects, only conditions that were found to possess added value at a 100% rate with all multidimensional data were reported. Results demonstrated that when the total sample was comprised of up to 30% of examinees with added value no more than 3% of the replications were found to possess added value when applying Haberman's (2008) method to the total sample. This result suggested that added value may not be invariant and points to the need to distinguish between individuals that may or may not possess data that would allow for reporting distinct and reliable raw subscores as a means of diagnostic information.

It should be noted that Sinharay and Haberman (2014) were the first in the literature to propose evaluating subscore added value invariance, but the purpose of their approach differed greatly from this study. Specifically, they suggested that subscore added value invariance should be conducted as a fairness evaluation for protected

minority groups to ensure that score interpretability is equivalent. In such an approach, if added value were found for one group and not for another, this would lead to follow-up analyses rather than differential reporting of subscores by group. Put simply by Sinharay and Haberman (2014), “In operational testing, one has either to report subscores for all the subgroups or not to report subscores for any subgroup” (p. 29). However, in this paper it was argued that the utility of reporting subscores for an individual should not be based on one’s manifest characteristics (e.g., gender or ethnicity), but rather on individual needs for diagnostic information, which is largely driven by a degree of multidimensional data at the individual-level. Furthermore, it was argued that if distinct and reliable subscores can be reported for an individual, then such information should be reported to assist in improving instruction and learning, regardless of the demographic characteristics of the examinee.

5.3 Aberrant Detection Procedures for Assessing Subscore Added Value Invariance

As it was argued that the invariance of subscore added value should be based on test performance rather than demographic variables, it was proposed that multivariate outlier and non-parametric person-fit statistics should be applied to individual-level data to identify aberrant score profiles and response patterns respectively due to multidimensionality. The multivariate outlier detection procedure applied to this study was the Mahalanobis Distance measure, which has been found to be an adequate method for identifying individual-level data that diverge from unidimensionality (Yuan, Fung, & Reise, 2004). Although there are numerous person-fit statistics that are based on classical test theory, item response theory, and structural equation modeling, the H^T index was applied in this study as it has shown promise in most accurately identifying a number of

different aberrant response behaviors (e.g., Karabatsos, 2003). Simulation analyses were applied to evaluate the adequacy of each procedure, which was defined in terms of Type I error, power, and subscore added value classifications.

Across conditions, Type I error was found to be maintained at acceptable rates at or below 5%; however, it should be noted that the Type I error rate of the H^T index increased as the subdomain test length decreased. Figure 7 demonstrates this dependency for conditions that differed solely in terms of subdomain test length. Specifically, we can see in the upper portion of Figure 7 that the purely multidimensional data for a test with subdomain test lengths of 10 items produced a distribution of the H^T index that ranged in value from -0.12 to 0.015, which indicated that a small proportion of multidimensional cases should be classified as non-aberrant. As a result, the density of the H^T index distribution for the combined data (both unidimensional and multidimensional data) was increased around the cut-value of 0, which led to increased Type I errors. In comparison, the lower portion of Figure 7 shows that for a test with subdomain test lengths of 50 the H^T index distribution for the multidimensional data were less variable and the values were predominately constricted between -0.1 and 0. As a result, the combined distribution was distinctively bimodal with the means of both modes being relatively distant from the cut-point. This resulted in decreased Type I error rates that were near 0%; however, it should be noted that even with short subdomain test lengths of 10 items, the Type I error rates were on average equal to 2%, which was mainly due to the distinctive bimodal distributions of the H^T index for the combined data across conditions. This distinctiveness also led to extremely high power rates that were near 100% across conditions and were found to be relatively independent of subdomain test lengths,

proportion of multidimensionality, and inter-subdomain correlations. Due to the low Type I error and high power rates that were observed, the H^T index was found to possess minimal bias in subscore added value classifications.

In contrast to the H^T index, the Mahalanobis Distance index was found to possess differential power rates by condition. That is, although the Type I error rates were maintained at 5%, power was found to be dependent on the proportion of multidimensional cases in the sample, subdomain test length, and inter-subdomain correlations. Specifically, power increased as the subdomain test lengths increased, while decreasing rates were observed as both the proportion of multidimensional cases and inter-subdomain correlations increased. This dependency is clearly illustrated in Figure 8, which shows the degree of overlap between the distributions of the generated unidimensional and multidimensional cases by condition. Although Yuan, Fung, and Reise (2004) proposed the use of Mahalanobis Distance in assessing unidimensionality, they only evaluated the method using applied datasets and as a result, did not assess Type I error and power rates. Consequently, there is no previous research in relation to Type I error and power rates for the Mahalanobis Distance measure that may support the findings of this study.

A closer examination of the aberrant cases identified by the Mahalanobis Distance measure demonstrated that only the more extreme multidimensional cases were identified. Such an assertion was supported by comparing the descriptive statistics of the generated multidimensional data and the aberrant cases identified by this method. In doing so, one sees that across conditions the aberrant cases tended to possess lower mean subdomain scores and greater variability (Table 10). To better understand why this

occurred, one can simply examine the item characteristic curves seen in Figure 9. As the aberrant cases were generated from a multidimensional extension of the three-parameter logistic model, there was greater variation in the probability of correctly responding to an item at the lower end of the theta continuum due to the pseudo-guessing parameter, while such variation decreased towards the upper end of the continuum. As a result, a ceiling effect was observed for high ability simulees, whereas due to possible guessing effects there was greater variation for low ability simulees. This variability led to score profiles that possessed more dispersion than the mean sample score profile, which was relatively flat, and as the dispersion in score profiles increased, the probability of being identified as an aberrant case by the Mahalanobis Distance measure also increased. Consequently, the aberrant cases possessed lower absolute subdomain means and greater subdomain variability when compared to the generated multidimensional data; however, it should be noted that the difference in subdomain means was practically negligible. A closer examination of the subdomain variability showed that on average the variability in score profiles observed for the aberrant cases was larger than would be expected by random error.

The increased variability of the aberrant cases led to increased subdomain reliability and decreased inter-subdomain correlations. As the inter-subdomain correlations and reliability are the two pieces of information that drive inferences based on Haberman's (2008) method, added value was obtained for nearly 100% of the replications when assessing the aberrant cases identified using the Mahalanobis Distance measure. This finding was supported in an analysis of an operational dataset in which only the cases identified using the Mahalanobis Distance measure were found to provide

added value, while the total sample and cases identified using the person-fit index were found to have no added value. Taken together, these results suggested that the Mahalanobis Distance measure may be an adequate procedure to identify score profiles that may possess meaningful variability.

5.4 Limitations and Directions for Future Research

Although this study illustrates a promising solution to providing valid and reliable diagnostic information to stakeholders, there are a number of limitations that must be discussed. For one, clearly when conducting simulation analyses, the generalizability of findings is limited to the particular context that is created. Though a concerted effort was made to include the most pertinent independent variables along with respective levels, it was impossible to include everything of importance. One area of research that was not covered in the present study was investigating the impact of the number of subdomain dimensions on identifying examinees that significantly diverged from unidimensionality. Although manipulating the number of subdomains generated would not have impacted the H^T index as misfit was assessed using item-level data, the power rates of the Mahalanobis Distance measure may have been more influenced. Specifically, as only dichotomous items were examined, raw subdomain scores were employed to compute the Mahalanobis Distance measure. As mentioned, this was done largely as the Mahalanobis Distance measure is known to exhibit odd behavior with non-normal data as the underlying assumption is that the data are continuous. Therefore, by holding the number of subdomains evaluated at four across both simulation and real data analyses, all analyses concerning the Mahalanobis Distance measure were based only on four independent variables.

As this measure evaluates distances from the centroid based on the mean, variance, and covariance of p variables, increasing the variability within and between the p variables will increase power. The assertion that variability is necessary for accurate identification of multivariate outliers is supported by results from this study, which demonstrated higher power rates when the number of items within each subdomain increased, regardless of inter-subdomain correlations and percentage of multidimensionality in the sample. Therefore, it would be of interest to evaluate power rates for the Mahalanobis Distance measure when a test is comprised of either dichotomous or polytomous items with less than four subdomains (and more than one). Such an analysis is of particular interest as Sinharay (2010) found that in a review of 25 operational tests 52% of tests reported raw scores for either two or three subdomains. The evaluation of bivariate outliers can be accomplished via graphical procedures, such as the bagplot approach (Rousseeuw, Ruts, & Tukey, 1999), but further research should evaluate power rates of the Mahalanobis Distance measure when only reporting three subdomains based on the subdomain test lengths included in this study.

In terms of the methodologies implemented in this study to identify aberrant cases, there were two limitations. The first limitation was that only two methods among a number of possible methodologies were employed. As an example, the H^T index is merely one of 36 or more person-fit indices currently in the literature. Although it is one of the best performing indices in previous research (Karabatsos, 2003) and was found to perform exceptionally well in identifying multidimensional cases, it is based on evaluating similarities in score patterns by assuming the Guttman scaling principle. In hindsight, such an approach is limited in two ways. First, as with any person-fit, an

aberrant case can be due to a number of possible issues, not related to solution-based behavior (e.g., random responding). Secondly, as the objective of identifying aberrant cases is to assist in providing *some* examinees with distinct and reliable subscores, the focus of the assessment should be at the subdomain-level as opposed to the item-level, which is not the case with most person-fit indices. As a result, procedures that input subdomain scores as independent variables appear to be of greatest interest.

To this end, the Mahalanobis Distance index was employed in this study and was found to identify cases with added value at near 100% rates. However, it should be noted that this measure is merely one of numerous procedures that can be viewed as exploratory profile analysis methods, which include cluster analysis, configural frequency analysis, and profile analysis via multidimensional scaling (PAMS; Ding, 2001). Furthermore, the Mahalanobis Distance index identifies aberrant cases dichotomously, which largely ignores the within-group variability, whereas other exploratory procedures, such as PAMS, provide continuous person profile indices. As a result, further research should evaluate the comparability of these various exploratory profile analysis methods in identifying cases with poor performance on one or more subdomains when compared to the remaining subdomains of interest.

An additional limitation associated with the methodologies employed to identify aberrant cases was the use of single critical values for both the H^T and Mahalanobis Distance indices. Such an approach was limited in that it assumed that the distribution underlying the test statistic of interest was independent of the proportion of multidimensionality, degree of multidimensionality, and subdomain test length of the generated and applied data. However, it should be noted that such an assumption was

only applied to the Mahalanobis Distance index as the H^T index does not base the classification of aberrant score patterns on a statistical test. Instead, this index applies the heuristic rule that a negative correlation between an examinee's score pattern with the score patterns of the remaining examinees is indicative of an aberrant case. This general rule was found to be an excellent cut-point for maintaining Type I error and increasing power to near 1.00 across conditions in this study. In contrast, the Mahalanobis Distance index was assumed to follow a Chi-square distribution with a critical value at an alpha-level of .05. Evidence to support the question of whether this distribution functioned independently from the independent variables included in this study can be seen in Figure 8. Specifically, the inclusion of up to 30% multidimensional data that ranged in subdomain inter-correlations from .30 to .70 and subdomain test lengths ranging from 10 to 50 items looked to have little impact on the assumed Chi-square distribution, particularly when looking at the area where the critical value was set. Clearly, the critical value could have been decreased to identify more multidimensional cases, but as seen in Figure 8 that would have increased Type I error rates. Such an approach was found to be undesirable as the main objective of this study was to find a methodology to identifying distinct and reliable subscores for *some* examinees, while avoiding the possible consequential validity issues associated with providing subscore information that lacked adequate distinctiveness and reliability. As mentioned, the consequential validity of supplying such information is the possibility of incorrect high-stakes decisions associated with poor subdomain performance (e.g., remedial instruction and negative teacher accountability ratings) and wasted resources of attempting to improve instruction and learning for an area of need when the need is actually lacking. As a result, the assumption

of constraining the Chi-square critical value to be equal at an alpha-level of .05 for the Mahalanobis Distance index across conditions was found to be tenable.

An additional limitation associated with this study was the sole use of Haberman's (2008) procedure as a criterion for assessing subscore added value. Although such an approach is popular in both research and practical contexts, it holds a number of limitations. For one, it assumes that the model underlying the data is a simple structure correlated-traits model. As a result, when applying this criterion to assessing subscore added value invariance, it was assumed that all outliers that diverged from unidimensionality possessed data that fit this model. However, it is possible that the assumption of a correlated-traits model may have been untenable. Therefore, it is suggested that for future analyses if sample sizes permit, one can apply exploratory dimensionality procedures on one-half of the data and cross-validate with a confirmatory approach using the remaining data. If a multidimensional (e.g., a bifactor or higher-order) model other than the correlated-traits model is found to provide improved fit for the outlier data, an interesting concern arises. That concern is whether Haberman's (2008) model provides accurate inferences related to subscore added value that is robust to violations of the dimensionality assumption, which clearly is a question that requires further research. To avoid the dimensionality assumption inherent in Haberman's (2008) procedure, a simple approach would be to report subscores for the cases that diverge from unidimensionality using a MIRT model that is found to best fit the data based on dimensionality assessments.

The second assumption underlying Haberman's (2008) procedure is that subdomain reliability is equivalent across all examinees. To be fair, Brennan (2012) as

well as Feinberg and Wainer (2014) have made the same assumption in their methods for assessing subscore added value. In combination with the use of subdomain correlations, assuming that reliability is consistent across all examinees has led both Brennan (2012) as well as Feinberg and Wainer (2014) to conclude that their procedures for assessing subscore added value provide identical inferences as Haberman's (2008) procedure. Regardless, the assumption of equivalent reliability across all examinees may be untenable as reliability may be conditional on examinee ability. As a result, there have been recent calls for the inclusion of conditional reliability estimates to assess both IRT model selection for estimating subscores (Bulut, 2013) as well as assessing subscore added value (Raymond & Feinberg, 2015).

The latter approach specifically assesses subscore added value by taking the proportion of individual-level score profile variability and the mean conditional reliability across subtests for an examinee. Raymond and Feinberg's (2015) approach differs from that of Haberman (2008), Brennan (2012), Feinberg and Wainer (2014), as well as the general procedure proposed in this study by assessing added value not for a group, but rather for an individual. As such, examinee-level differences in subscore profile variability and score precision can be taken into consideration when evaluating added value. However, it should be noted that Raymond and Feinberg's (2015) approach is very much in the early stages of research. For example, there currently are no guidelines for making classifications of added value based on either hypothesis testing or heuristics. Clearly, additional research is needed to identify a threshold that may indicate meaningful variability. However, once a sensible criterion has been developed, future

research should look into the comparability of added value classifications between the approaches suggested in this paper and by Raymond and Feinberg (2015).

Lastly, this study raises practical concerns related to score reporting. Specifically, the concept that subscores may be valid for *some* examinees elicits an important question, how should differential examinee score information be reported to stakeholders? To the best of the author's knowledge, there has been little attention given to this area in the literature. Up to this point, researchers have suggested personalizing score reports (Goodman & Hambleton, 2004) and being cognizant of the intended audience's characteristics for improved score report design (Zapata-Rivera & Katz, 2014). As a result, further research is needed to better understand if stakeholders will be open to the idea that some examinees will receive diagnostic information, while others will not. For example, will teachers that perceive diagnostic information to be helpful appreciate the fact that not all students will receive such information because of measurement concerns? It is conceivable to believe that a parent could receive subdomain information for one child and not another. Would such differences in score reports cause confusion and ultimately, lead to a loss of confidence regarding assessment results? Although such uncertainty may have largely been caused by the current practice of providing diagnostic information to all examinees, it is hypothesized that a shift in perspective on subscore reporting will take both time and effort in explaining measurement concerns to stakeholders (e.g., Zwick, Zapata-Rivera, & Hegarty, 2014).

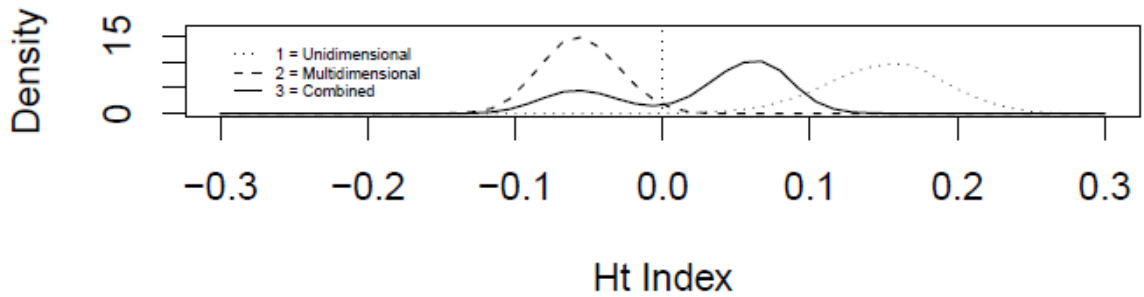
5.5 Conclusion

The results of this study have a number of important implications. For one, this study demonstrated the need to assess subscore added value invariance based on test

performance. Specifically, it was shown that up to 30% of examinees with added value can be masked when only evaluating the total sample. Secondly, the Mahalanobis Distance measure, which is a multivariate outlier detection procedure was found to show promise for identifying aberrant cases that possessed subscores of added value. Such an assertion was supported via both simulated and applied datasets. In particular, the Mahalanobis Distance measure was found to possess adequate Type I error rates (i.e., falsely identifying unidimensional score profiles as aberrant) and although it was found to possess lower power rates than the person-fit procedure (H^T index) included in this study, the cases identified most likely possessed high multidimensionality. As a result, these cases were found to have added value of nearly 100% when assessed as a group, regardless of generating subdomain test-lengths or inter-subdomain correlations. In contrast, the H^T index identified cases that possessed added value only when subdomain test lengths were comprised of 25 or more items and moderate inter-subdomain correlations. To support these findings, a large-scale dataset was analyzed and of the two procedures only the Mahalanobis Distance measure was found to provide added value for about 7% of the sample when no added value was obtained for the total sample or the cases identified using the H^T index. Closer examination of the aberrant cases identified by the Mahalanobis Distance measure for both simulated and operational datasets showed that the average subscore profile was more variable than would be expected based on random error. This result supports the idea that the Mahalanobis Distance measure is able to identify cases with meaningful variability that may allow for both valid and reliable subdomain inferences.

Besides the methodological implications that this study provides, it also sheds light on a new perspective on subscore reporting, which is that *subscores are not for everyone*. Traditionally, from a psychometrician's perspective, the decision to provide subscore information is evaluated for the total sample. Such a perspective unnecessarily frames subscore reporting dichotomously as either being useful or not. However, framing the question of subscore utility in this manner ignores one simple truth, subscores may be informative for some individuals within the total sample, whereas they may be uninformative for others. This study proposed an approach that could be sensitive to this lack of invariance and is practical in a number of ways. For one, it does not require either multidimensional modeling or overhauling the test development process as has been suggested by some researchers (e.g., Wainer, Sheehan, & Wang, 2000). Secondly, it can easily be applied in operational testing programs as the calculations can be quickly conducted in Excel or basic general statistical software packages, such as SPSS, SAS, or R. Taken together, this approach and shift in perspective concerning subscore reporting may allow testing programs to meet both legislative and stakeholder demands for diagnostic information, while also ensuring that the subscores provided to *some* examinees are of adequate psychometric quality, which may allow for valid subdomain inferences.

I.Sub=10, r=.70, 30% Multidim



I.Sub=25, r=.70, 30% Multidim

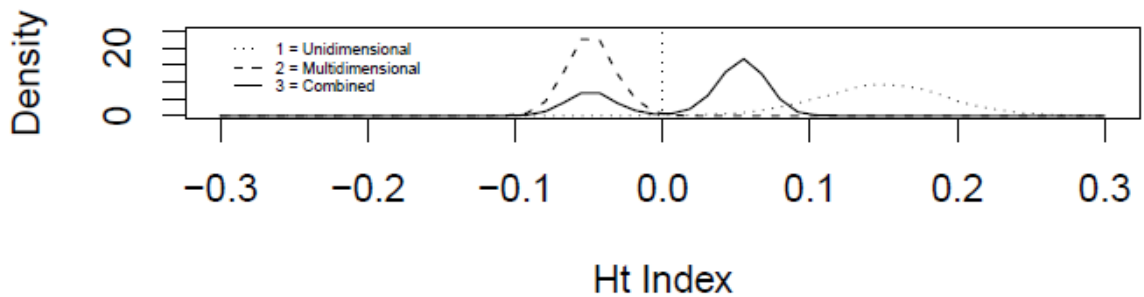


Figure 7. H^T distributions by generating dimensionality and condition. I.Sub = the number of items by subdomain; r = inter-subdomain correlations of the generating multidimensional data; Multidim = the percentage of multidimensional data in the combined dataset. The vertical dotted line denotes the critical value employed to classify aberrant cases.

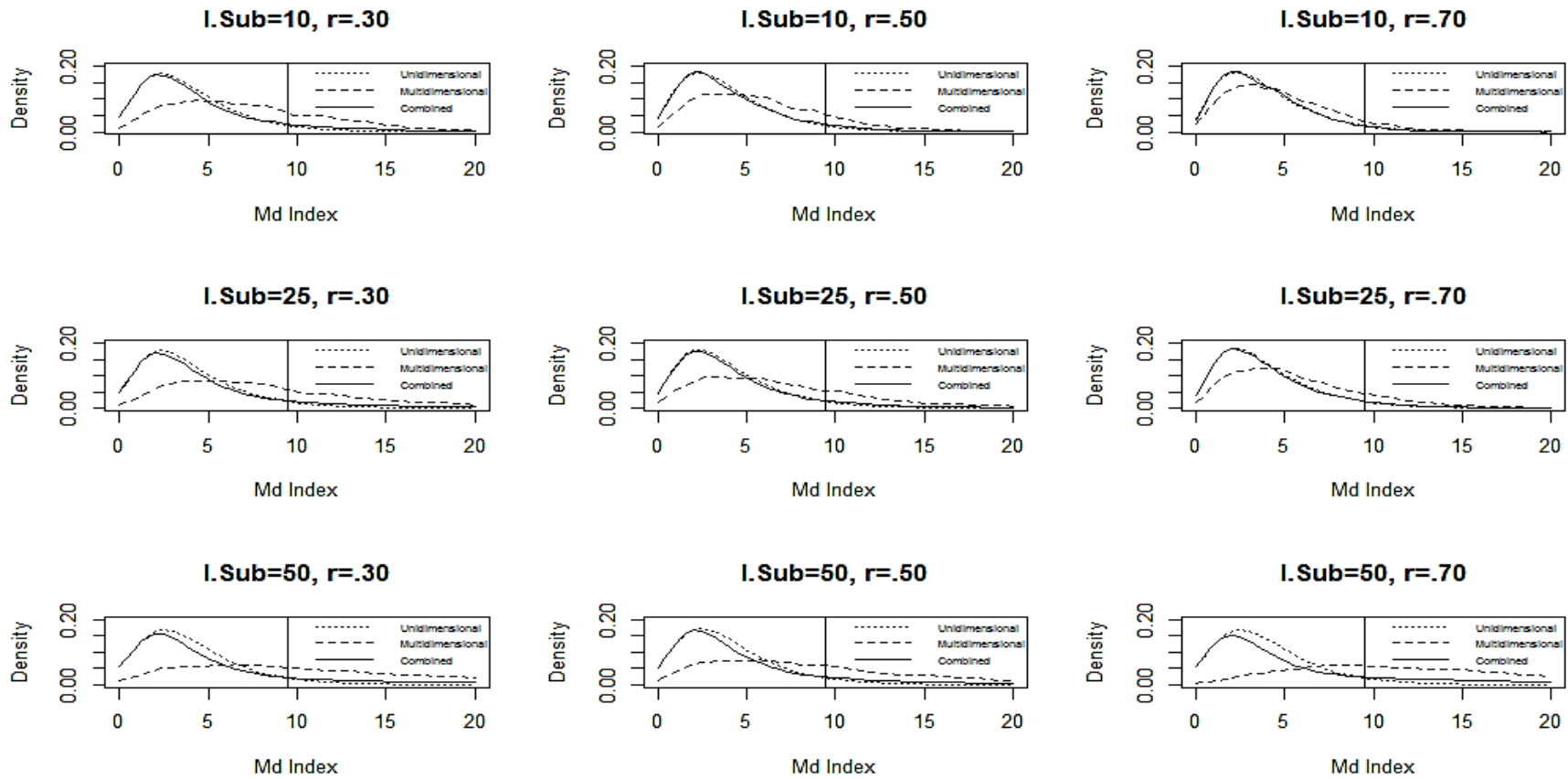


Figure 8. Distributions of the Mahalanobis Distance index by generating dimensionality and condition. I.Sub = the number of items by subdomain; r = inter-subdomain correlations of the generating multidimensional data; the percentage of multidimensional data in the combined dataset was constrained to 30%. The vertical dotted line denotes the critical value implemented for classifying aberrant cases.

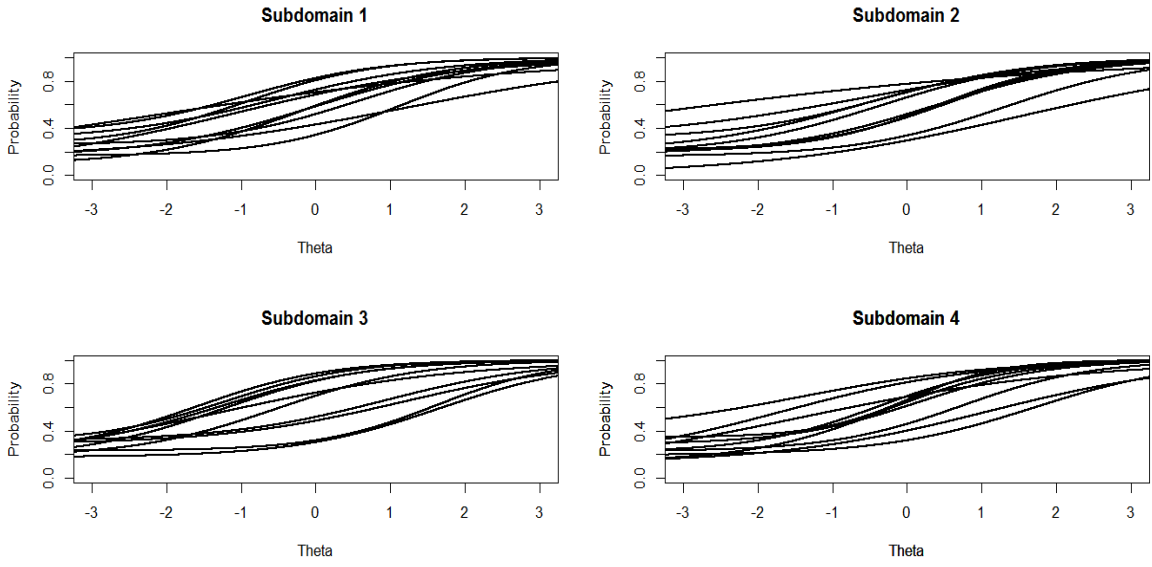


Figure 9. Item characteristic curves for each item by subdomain with a subdomain test length of 10 items.

REFERENCES

- Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York, NY: Dover.
- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*(2), 113-127.
- Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive measurement of multiple abilities*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1-23.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association.
- Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the lz person-fit statistic. *Practical Assessment, Research & Evaluation*, *12*(16).
- Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement*, *70*(1), 22-35.

- Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4), 541-561.
- Ben-Gal, I. (2005). Outlier detection. In O. Maimon and L. Rockach (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers*. New York, NY: Springer.
- Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34(3), 197-211.
- Brennan, R. L. (2012). *Utility indexes for decisions about subscores* [CASMA Research Report 33]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.
- Buluth, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.
- Capizzi, A. M., & Fuchs, L. S. (2005). Effects of curriculum-based measurement with and without diagnostic feedback on teacher planning. *Remedial and Special Education*, 26(3), 159-174.
- Caussinus, H., & Ruiz, A. (1990, January). Interesting projections of multidimensional data by means of generalized principal component analyses. In K. Momirovic and V. Mildner (Eds.), *Proceedings from Computational Statistics, 9th Symposium held at Dubrovnik, Yugoslavia*. New York, NY: Springer.
- Chalmers, P. (2014). *mirt: Multidimensional Item Response Theory*. R package version 1.5, URL <https://github.com/philchalmers/mirt>

- Coe, R. (1998). Can feedback improve teaching? A review of the social science literature with a view to identifying the conditions under which giving feedback to teachers will result in improved performance. *Research Papers in Education*, 13 (1), 43-66.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
- Davey, T., & Hirsh, T. M. (1991, April). *Examinee discrimination as measurement properties of multidimensional tests*. Annual meeting of the National Council on Measurement in Education, Chicago, IL.
- de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30(3), 295-311.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement* 33(8), 620-639.
- de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, 35(4), 296-316.
- Ding, C.S. (2001). Profile analysis: Multidimensional scaling approach. *Practical Assessment, Research, & Evaluation*, 7(16). Retrieved from <http://pareonline.net/getvn.asp?v=7&n=16>

- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*(1), 59-79.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86.
- Edwards, M.C, & Vevea, J.L. (2006). An empirical bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241-259.
- Elawar, M.C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77*(2), 162-173.
- Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186.
- Feinberg, R.A., & Wainer, H. (2014). A simple equation to predict a subscore's value. *Educational Measurement: Issues and Practice, 33*, 55-56.
- Faulkner-Bond, M., Shin, M., Wang, X., Sireci, S.G., & Zenisky, A.L. (2013, April). *Score reports for English proficiency assessments: Current practices and future directions*. Paper presented at the Annual Conference of the National Council on Measurement in Education, San Francisco, CA.
- Ferrando, P.J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42*(3), 481-507.

- Firestone, W.A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*, 43(2), 100-107.
- Fu, J., & Li, Y. (2007, April). *An integrative review of cognitively diagnostic psychometric models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Fu, J., Qu, Y. (2012, April). *A review of subscore estimation methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
- Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17(2), 145-220.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.
- Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psychology*, 66(3), 452-469.
- Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, 62(1), 79-95.
- Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (RR-08-45). Princeton, NJ: Educational Testing Service.

- Haidi A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, 54, 761-771.
- Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on the comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Hambleton, R.K. & Zenisky, A.L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K.F. Geisinger, B.A. Bracken, J.F. Carlson, J.C. Hansen, N.R. Kuncel, S.P. Reise, & M.C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology (Volume 3)*. Washington, D.C.: American Psychological Association.
- Hardin, J., & Rojke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44(4), 625-638.
- Hauser, R. M., & Heubert, J. P. (1998). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, D.C.: National Academies Press.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

- Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied psychological measurement*, 28(6), 407-426.
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15(1), 136-153.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298.
- Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.
- Kronenfeld, D.B. (1972). Guttman scaling: Problems of conceptual domain, unidimensionality, and historical inference. *Man*, 7(2), 255-276.
- Kunnan, A. J. and Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 610–625). Walden, MA: Wiley-Blackwell.
- Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, 6(3), 172-189.
- Li, M. N. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21(3), 215-231.
- Linacre, J.M. (2012). A comment on the H^T person fit statistic. *Rasch Measurement Transactions*, 26(1), 1358.
- Livingston, S. A., & Rupp, S. L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning teachers* (Research Report 04-48). Princeton, NJ: Educational Testing Service.

- Luecht, R.M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (April, 2006). *Scalability and the Development of Useful Diagnostic Scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Lyrén, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, & Evaluation, 14*(4), 1-10.
- Marsh, J.A., Pane, J.F., & Hamilton, S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica, CA: RAND. Massachusetts Department of Education (2013). *2013 MCAS and MCAS-ALT Technical Report*. Malden, Massachusetts: Author.
- McDonald, R. P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York, NY: Springer.
- Md Desa, Z.N.D. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*(1), 3-8.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107-135.

- Muraki, E. & Carlson, J.E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19(1), 73-90.
- Muralidharan, K., & Sundararaman, V. (2010). The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India. *The Economic Journal*, 120(546), F187-F203.
- Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide [Seventh Edition]. Los Angeles, CA: Muthén & Muthén
- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, 22(1), 53-69.
- Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research & Evaluation*, 8. Retrieved from <http://pareonline.net/getvn.asp?v=8&n=11>
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation*, 9(6), 1-12.
- Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: cases and their influence. *Multivariate Behavioral Research*, 46(2), 202-228.
- Puhan, G., Sinharay, S., Haberman, S. & Larkin, K. (2008, April). *Comparison of subscores based on classical test theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Raymond, M., & Feinberg, R. (2015, April). *Subscores aren't for everyone: Alternative strategies for evaluating subscore utility*. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago, IL.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.
- Reise, S.P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.
- Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*(1), 3-21.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.
- Rizopoulos, D. (2014). *Latent trait models under IRT*. R package version 1.0-0. URL <http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm>
- Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*(1), 47-57.
- Rosenbaum, P.R. (2009). *Design of observational studies*. New York, NY: Springer.

- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388), 871-880.
- Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223.
- Rousseeuw, P. J., Ruts, I., & Tukey, J.W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, 53, 382-387.
- Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4), 219-262.
- Shin, D. (2007). *A Comparison of Methods of Estimating Subscale Scores for Mixed-Format Tests*. Austin, TX: Pearson.
- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7(22), 131-145.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16(2), 149-157.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150-174.
- Sinharay, S. (2013). A Note on Assessing the Added Value of Subscores. *Educational Measurement: Issues and Practice*, 32(4), 38-42.

- Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement, 51*(2), 212-222.
- Sinharay, S., & Haberman, S. J. (2014). An empirical investigation of population invariance in the value of subscores. *International Journal of Testing, 14*(1), 22-48.
- Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.
- Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45*(3), 553-573.
- Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40.
- Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357-375.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*(1), 33-66.
- Stein, M. (2003). *Making sense of the data: Overview of the K-12 data management and analysis market*. Retrieved August 10, 2014, from http://www.eduventures.com/about/press_room/11_18_03.cfm

- Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63-86.
- Stricker, L. J. (1993). *Discrepant LSAT subscores* (Technical Report No. 93-01). Newtown, PA: Law School Admission Council.
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Tao, S. (2009). *Using collateral information in the estimation of sub-scores: A fully Bayesian approach* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.
- Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied measurement in education, 17*(2), 89-112.
- Tendeiro, J.N. (2014). *Person fit*. R package version 1.2. URL <http://cran.r-project.org/web/packages/PerFit/PerFit.pdf>
- Tendeiro, J.N., & Meijer, R.R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*(3), 239-259.
- Trout, D. L., & Hyde, E. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

- Tyler, J.H. (2013). If you build it will they come? Teachers' online use of student performance data. *Education Finance and Policy*, 8(2), 168-207.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287-307.
- Wainer, H., Sheehan, K.M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, 37(2), 113-140.
- Wainer, H., Vevea, J.L., Camacho, F., Reeve, B.B., Rosa, K., Nelson, L., et al. (2001). Augmented scores: "Borrowing Strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wang, J., Zamar, R., Marazzi, A., Yohai, V., Barrera, M., Maronna, R.,...Konis, K. (2014). *Robust Statistical Methods*. R package version 0.4-16.
- Wang, X., Faulkner-Bond, M., & Shin, M. (2012, October). *Subscore reporting in state K-12 content assessments: Review of current practices*. Paper presented at the Conference, Amherst, MA.
- Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45(4), 479-494.
- Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31(2), 83-105.
- Yen, W. M. (1987, June). A Bayesian/IRT index of objective performance. *Annual meeting of the Psychometric Society, Montreal, Quebec, Canada*.

Zapata-Rivera, J.D., & Katz, I.R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, 4, 442-463.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19, 116-138.