Reliability and Validity Evidence of Diagnostic Methods: Comparison of Diagnostic Classification Models and Item Response Theory-Based Methods

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA

BY

Yoo Jeong Jang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Michael C. Rodriguez, Adviser

August, 2022

Acknowledgments

Foremost, I would like to acknowledge and give my deepest thanks to my advisor, Dr. Michael C. Rodriguez, for accepting me to the QME program and patiently guiding me through this long journey with his invaluable expertise and support. Without his advice and encouragement, this dissertation would not have been completed. I am very proud of and grateful for being his advisee.

I also would like to express my gratitude to my defense committee, Dr. Mark L. Davison, Dr. David R. Johnson, and Dr. Joseph Rios, who generously shared their knowledge and insights. Their valuable feedback helped me stay on the right track throughout the dissertation process.

I am also grateful to my friends and colleagues in the QME program. They inspired me with their sincerity and enthusiasm for the study of educational and psychological measurement. I also would like to extend my sincere thanks to the staff, especially Lori Boucher, in the QME program. Their continuous assistance and patient guidance helped me complete academic milestones during my graduate studies.

Words cannot express my gratitude to my parents and siblings, Hyoseok and Yoonjeong. Their love and belief in me have kept my spirits and motivation high during this process. Thank you for being my unwavering advocate.

I am also deeply indebted to my husband, Seungsuk, who has been a great listener and shared his insights as an experienced researcher during my graduate studies. Thank you for seeing the best in me and encouraging me whenever I doubted myself.

Lastly, my precious children, Brian and James, have been the greatest joy and comfort during this whole process and have given me a strong sense of purpose. I hope my accomplishment can serve as a reminder to them that they can achieve what they want if they do not give up and keep trying.

Dedication

This dissertation is dedicated to my parents, Ki-Bok Jang and Soon-Ae Do, whose unconditional love, prayers, and tremendous support carried me through this whole journey. Everything I have and everything I am, I owe it all to them.

Abstract

Despite the increasing demand for diagnostic information, observed subscores have been often reported to lack adequate psychometric qualities such as reliability, distinctiveness, and validity. Therefore, several statistical techniques based on CTT and IRT frameworks have been proposed to improve the quality of subscores. More recently, DCM has also attracted increasing attention as a powerful diagnostic tool that can provide fine-tuned diagnostic feedback. Despite its potential, there has been a dearth of research evaluating the psychometric quality of DCM, especially in comparison with diagnostic methods from other psychometric frameworks. Therefore, in this simulation study, DCM was compared with two IRT-based subscore estimation methods in terms of classification accuracy, distinctiveness, and incremental criterion-related validity evidence of subscores. Manipulated factors included diagnostic methods, subscale length, item difficulty distribution, intercorrelations of subscores, and criterion validity coefficients. For classification accuracy, all diagnostic methods yielded comparable results when the center of item difficulty coincided with mean examinee ability and cut-scores. However, when average item difficulty was mismatched with mean examinee ability and cut-scores, DCM yielded substantially higher/lower classification accuracy than IRT-based methods with direction and magnitude of discrepancy depending on the type of agreement measures employed. For subscore distinctiveness, compared to IRT-based methods, DCM yielded subscores more distinct from each other and overall scores when continuous rather than discrete subscores were utilized. Lastly, regarding incremental criterion-related validity evidence, the contribution of DCM estimates over and above overall scores tended to be comparable to but slightly smaller than that of IRT-based methods. Additionally, higher classification accuracy was associated with longer subscales, item difficulty distribution more aligned with examinee ability distribution and cut-scores, and higher intercorrelations of subscores. The same conditions except for higher

intercorrelations of subscores also tended to be associated with higher subscore distinctiveness. In contrast, incremental criterion-related validity evidence of subscores was largely a function of intercorrelations of subscores and magnitude of criterion validity coefficients: it increased with lower intercorrelations of subscores and higher criterion validity coefficients. In general, the results of this study suggested that IRT-based methods would be preferable over DCM as diagnostic means when item responses are obtained from IRT-based assessment forms.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Background

Educational stakeholders have become increasingly interested in obtaining scores on the subscales for diagnostic purposes (Brennan, 2012; Sinharay et al., 2011). Students want to get detailed information about their strengths and weaknesses to remediate the latter. Similarly, teachers want to utilize such information for making placement decisions and providing instruction customized to the student's individual needs (Firestone, 2014; Haladyna & Kramer, 2004; Kunnan & Jang, 2009). Despite their potential to provide diagnostic information, as with the total score, subscores must meet professional quality standards before being reported to stakeholders (Haberman, 2008; Tate, 2004). This is in line with Standards 1.14 and 2.3 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) which require that each reported score must provide sufficient evidence of reliability, distinctiveness, and validity for the intended purposes.

Subscore reliability is especially important when the primary goal of the test is to provide diagnostic information (Feinberg & Wainer, 2014). According to Sinharay et al. (2019), an adequate level of reliability depends on which types of comparisons and decisions will be made according to subscores. For subscores to be of any remedial use, minimum reliability of .8 or higher is deemed adequate (Nunally, 1978; Sinharay et al., 2019), whereas minimum reliability of .7 is in general required for test scores (Choi & Papageorgiou, 2020). However, subscores were often found to have low reliability mainly because subscores are typically based on a smaller number of items compared to the total score. According to a comprehensive review of score reporting practices by Goodman and Hambleton (2004), some testing programs computed

subscores based on as few as five items, whereas researchers (e.g., Sinharay, 2010) of simulation studies have indicated that at least 20 multiple-choice items are needed per each subscale for an adequate level of subscore reliability. Operational constraints such as time limits, test fatigue (especially for young learners), and costs to develop and administer items often resulted in short subscales which lacked adequate information about subdomains of interest (Sinharay et al., 2019).

Subscore distinctiveness is divergent evidence for validity. As the correlation between the subscore and the rest of the test gets smaller, the subscore is more likely to add unique information. Researchers (e.g., Haberman, 2008; Sinharay, 2010; Sinharay et al., 2007), however, often found that subscores were highly correlated with each other. High intercorrelations among subscores are mainly attributable to retrofitting approach prevalent in educational testing in which subscores are obtained from essentially unidimensional assessments that were not designed to provide diagnostic information on examinees (Haberman, 2008; Luecht et al., 2006; Sinharay et al., 2019).

Likewise, incremental validity or added value of subscores has to do with the extent to which the smaller subset of items used for subscores provides new information about examinee abilities that is not already provided by the total test score (Haberman, 2008). Researchers (e.g., Haberman, 2008; Haberman et al., 2009; Puhan et al., 2010; Sinharay, 2010) indicated that subscores are more likely to add value over and above total scores if they are highly reliable and if they are distinct from the rest of the test scores. As such, the distinctive contribution of each subscore can constitute meaningful signals rather than noise only when subscores are accurate and stable enough (Choi & Papageorgiou, 2020).

Subscore Estimation Methods from CTT and IRT Frameworks

Given the lack of subscore reliability and distinctiveness and hence the subscore added value, several statistical techniques to improve the psychometric quality of subscores have been proposed from classical test theory (CTT) and item response theory (IRT) frameworks. The common thread running through these methods is that they utilize collateral or ancillary information to improve the subscore estimation (e.g., de la Torre, 2008b; de la Torre & Patz, 2005; Haberman, 2008; Haberman & Sinharay, 2010; Kahraman & Kamata, 2004; Wainer et al., 2001; W.-C. Wang et al., 2004; Yao & Boughton, 2007; Yen, 1987). The ancillary information typically involved in these methods is the examinee's performances on items outside the target subscale (i.e., out-of-scale items). According to de la Torre et al. (2011), this type of information is considered in-test collateral information because it is inherent in the test and can be obtained with the examinees' item responses. In contrast, out-of-test collateral information includes demographic variables (e.g., sex, age, and race) and education variables (e.g., grade level and courses taken) that are not contained in item responses but provide additional information about examinee characteristics (de la Torre et al., 2011).

Despite the commonality, these approaches using ancillary information differ in terms of how they utilize the ancillary information to estimate subscores. Some approaches (e.g., Haberman, 2008; Wainer et al., 2001) use information from the examinee's performance on other subdomains or on the overall test to augment the target subscore. Alternatively, some other techniques (e.g., de la Torre, 2008b; de la Torre & Patz, 2005; Haberman & Sinharay, 2010; W.-C. Wang et al., 2004; Yao & Boughton, 2007) explicitly models the multidimensionality presumed to underlie the assessment and then incorporate the correlational structure of the examinee's subscale performances in the estimation procedure to simultaneously estimate subscores. These correlation-based methods result in estimates that are more accurate than raw subscores or non-augmented ability estimates calibrated from a unidimensional model. However,

researchers have demonstrated that these methods improved the subscore reliability at the expense of decreasing the subscore distinctiveness (Feinberg, 2012; Feinberg & Wainer, 2014; Stone et al., 2010).

Diagnostic Classification Models

Aside from CTT and IRT frameworks, diagnostic classification models (DCMs), also known as cognitive diagnostic models (CDMs), have attracted increasing attention as a means to provide individual examinees with detailed diagnostic information as to multiple latent traits underlying the test. DCMs are a special case of latent class models (LCMs) wherein examinees are classified into discrete latent classes based on mastery of multiple latent traits, also referred to as skills or attributes (Rupp & Templin, 2008b). Therefore, in contrast to typical educational test score reports presenting a small number of content-based subscores, DCMs provide an attribute profile indicating whether the examinee meets the criteria for diagnosis on one or more of the latent traits, hence, informing the examinee's specific strengths and weaknesses (Roberts & Gierl, 2010).

Owing to the assumption of discrete rather than continuous latent traits, DCMs can achieve the same level of reliability as IRT with fewer data demands while allowing for the modeling of more complex data structures (Templin & Bradshaw, 2013). However, in contrast to the abundance of literature on the quality of subscores estimated from the CTT or IRT framework, there has been a paucity of researchers evaluating the reliability, validity, and utility of DCM results (Bradshaw et al., 2014; Haberman & von Davier, 2007; Jurich & Bradshaw, 2014; Rupp & Templin, 2008b; Sessoms & Henson, 2018; Sinharay & Haberman, 2009). In an extensive literature review of DCM applications, Sessoms and Henson (2018) noted that of 36 applied DCM research conducted in 2009 or later, only 36% and 22% reported the reliability and validity evidence of diagnostic feedback provided by DCMs, respectively.

Statement of the Problem

Given that the quality of subscores exerts a substantial impact on inferences about examinees' strengths and weaknesses which in turn inform further remedial and instructional effort, it is crucial to establish the psychometric quality of subscores. However, observed subscores have been often reported to have low reliability and/or high correlations with the rest of the test. Although several statistical techniques have been successful in improving the reliability of subscore estimates, they are still limited in that they inevitably decrease the subscore distinctiveness.

Researchers have indicated that DCMs may be a viable alternative to extant subscore estimation methods in that DCMs can provide detailed diagnostic feedback with greater accuracy. Despite the theoretical appeal, applications of DCMs have been often criticized for lacking reliability and validity evidence and thus rarely guiding practical decision-making (Sessoms & Henson, 2018). Moreover, given the complexity involved in the estimation and interpretation of DCMs, researchers

(e.g., Montero et al., 2003; Sinharay & Haberman, 2009; Wilhelm & Robitzsch, 2009) further called for proper justification for selecting DCMs over simpler models from CTT or IRT framework. At present, however, there is a lack of studies comparing DCMs and other diagnostic methods in a systematic fashion with respect to reliability, distinctiveness, and validity of subscores. Therefore, it is warranted to evaluate the psychometric quality of DCM estimates under various statistical conditions, especially in comparison with subscore estimates from diagnostic methods based on different psychometric frameworks.

Purpose of the Study

To address the gap in the literature, the present study had three primary purposes. First, I aimed to compare DCM and other diagnostic methods in terms of subscore reliability under

various conditions. Second, I aimed to compare DCM and other diagnostic methods in terms of subscore distinctiveness under various conditions. Third, I aimed to compare DCM and other diagnostic methods in terms of subscore added value under various conditions. Results of this study will provide reliability, distinctiveness, and validity evidence of DCM relative to those of other subscore estimation methods. Therefore, findings from this study will help ensure that diagnostic information provided by DCMs leads to accurate decisions about examinees as well as improvement in instruction and student learning. Findings from this study will also inform researchers and practitioners of the relative strengths and limitations of different diagnostic methods under various conditions, thus guiding their selection of the diagnostic method that is most suitable for the intended use of test scores.

Chapter 2

Literature Review

In this chapter, literature on methods for evaluating the psychometric quality of subscores and estimating subscores is reviewed. This chapter is organized into the following four sections:

1. Techniques to Evaluate Technical Adequacy of Subscores. This section includes techniques to evaluate the distinctiveness and added value of subscores. Especially, three statistical techniques to evaluate the added value or incremental validity evidence of subscores are described in greater detail: (a) Haberman's (2008) added value analysis, (b) between-person and within-person subscore reliability with profile analysis (Bulut et al., 2016), and (c) criterion-related validity evidence for subscores (Davison et al., 2015).
2. Subscore Estimation Methods from CTT and IRT Frameworks. In this section, CTT- and IRT-based subscore estimation methods that are most widely used in research and practice are reviewed focusing on their theoretical background and strengths and limitations. Three methods reviewed in this section include (a) regression approach such as Wainer et al.'s (2001) subscore augmentation and Haberman's (2008) weighted averages, (b) subscore augmentation with theta estimates from IRT models, and (c) multidimensional item response theory (MIRT) approach. The detailed description of these methods is followed by a review of studies where researchers have compared the technical adequacy of these approaches under various conditions.
3. Diagnostic Classification Models. This section contains a comprehensive review of DCMs comprising five subsections. The first section illuminates definitional properties of DCMs that are distinct from other diagnostic methods reviewed in the previous section. The next three subsections illustrate the psychometric properties of

DCMs focusing on (a) attribute specification, (b) psychometric models of DCMs and model selection procedure, and (c) quality of DCM estimates. They are followed by the fifth subsection reviewing the advantages and limitations of DCMs in comparison to other subscore estimation methods.

4. Summary of Literature Review and Research Questions. This section includes summary of findings reviewed in the previous sections and presents a set of research questions synthesized to address the gap in the literature regarding subscore estimation.

It is noteworthy that latent traits, subdomains, attributes, and skills are treated interchangeably for this study although in the context of DCMs, skills and attributes refer to unobserved dichotomous properties underlying the behaviors (Rupp et al., 2010).

Techniques to Evaluate Technical Adequacy of Subscores

Although all methods for evaluating the quality of subscores are concerned with the variance specifically attributed to subscores (Choi & Papageorgiou, 2020), they differ in terms of their main focus (Sinharay et al., 2019): subscore distinctiveness (i.e., dimensionality or internal structure of a test) or subscore added value (i.e., incremental validity evidence).

Techniques to Evaluate Subscore Distinctiveness

Procedures focusing on the subscore distinctiveness evaluate the extent to which a multidimensional model can explain the covariance among items better than a unidimensional model (Biancarosa et al., 2019; Choi & Papageorgiou, 2020). These techniques include disattenuated inter-subdomain correlations (e.g., Lyrén, 2009; McPeek et al., 1976; Sinharay, 2010), factor analytic approaches such as principal component analysis (PCA), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) (e.g., Sinharay et al., 2007; Stone et al., 2010), the beta-binomial model (Harris & Hanson, 1991), multidimensional item response theory

(MIRT) models (e.g., Ackerman et al., 2003; Reckase, 1997; von Davier, 2008), and formal dimensionality tests such as DIMTEST (Stout, 1987) and DETECT (Zhang & Stout, 1999). Readers interested in more details about these procedures are referred to Sinharay et al. (2019).

Techniques to Evaluate Added Value Evidence for Subscores

Procedures focusing on the added value or incremental validity evidence of subscores take into consideration both reliability and distinctiveness of subscores; thereby, they evaluate the degree to which all accumulated evidence supports the interpretation of the test scores for their intended use (Sinharay et al., 2019). In what follows, three procedures in this category are described.

Haberman's Added Value Analysis. Haberman (2008) proposed a CTT-based method evaluating the extent to which the observed subscore can predict the true subscore more accurately than can the observed total score. To apply this method, the true subscore (T_{jk}) of examinee *j* on subtest *k* is estimated in the following two ways:

1. The true subscore is estimated based on the observed subscore (X_{jk}) of examinee *j* on subtest *k* as follows:

$$\hat{T}_{jk} = \bar{X}_k + a\left(X_{jk} - \bar{X}_k\right), \quad (1)$$

where $\bar{X}_k$ is the mean subscore on subtest *k* for the sample of examinees, and a is the reliability of subtest *k* which is obtained by coefficient alpha.

2. The true subscore is estimated based on the observed total score (X_j) of examinee *j* as follows:

$$\hat{T}_{jk} = \bar{X}_k + c\left(X_j - \bar{X}\right), \quad (2)$$

where X_j is an observed total score for examinee j, $\bar{X}$ is the mean total score for the sample of examinees, and c is constant based on the reliabilities and standard deviations of the observed subscores and total score as well as the correlations among the observed subscores.

The quality of each subscore estimate can be evaluated using the mean squared error (MSE) which is defined as the average squared difference between subscore estimates and true subscores as follows:

$$MSE = E\left[T_{jk} - \hat{T}_{jk}\right]^2, \tag{3}$$

where $\hat{T}_{jk}$ is obtained by adopting one of the subscore estimates described above. The smaller the MSE, the more accurate the subscore estimate. To quantify how much of the MSE can be reduced by adopting each subscore estimate (i.e., Equation 1 or Equation 2) in comparison to utilizing the mean subscore (i.e., $\hat{T}_{jk} = \bar{X}_k$) as the constant predictor of the true subscore, the proportional reduction in MSE (PRMSE) can be computed for each subscore estimate as follows:

$$PRMSE = \frac{MSE\ for\ the\ constant\ predictor - MSE\ for\ a\ certain\ subscore\ estimate}{MSE\ for\ the\ constant\ predictor}. \tag{4}$$

The PRMSE for the observed subscore as a predictor of the true subscore ($PRMSE_s$) has been demonstrated to be equal to subscore reliability (Sinharay et al., 2011). In contrast, the PRMSE for the observed total score as a predictor of the true subscore ($PRMSE_x$) is represented by

$$PRMSE_x = \rho^2(T_k, T)\rho^2(T, X), \tag{5}$$

where T_k is the true subscore for subtest k, T is the true total score, X is the observed total score, $\rho^2(T, X)$ is equal to the total test score reliability, and $\rho^2(T_k, T)$ can be computed using the sample variances, reliabilities, and correlations (i.e., disattenuated ones corrected for unreliability) of the total score and the subscores (Sinharay & Haberman, 2008).

The PRMSE is a reliability-like statistic and typically ranges from 0 to 1 with a higher value indicating a greater degree of trustworthiness. In the context of subscore estimation, the larger the PRMSE, the more accurate the subscore estimate because a larger PRMSE corresponds to a smaller MSE in estimating the true subscore (Haberman, 2008). Thereby, the observed subscores are deemed to provide the added value over and above the observed total score if $PRMSE_s$ is greater than $PRMSE_x$. As stated earlier, this is more likely if subscores have high reliability and if subscores are distinct from the rest of the test scores (Haberman, 2008; Haberman et al., 2009; Puhan et al., 2010; Sinharay, 2010).

Owing to the straightforward implementation and the concrete decision rule, the value-added analysis of Haberman (2008) has been popular in both research and practice (Feinberg & Wainer, 2014; Puhan et al., 2010; Sinharay, 2010; X. Wang et al., 2019). Its use has also been extended to MIRT models (Haberman & Sinharay, 2010; Thissen, 2013) although PRMSE was first applied in the context of CTT. However, it should be noted that this method assumes that data is based on a multidimensional simple structure correlated-traits model. Therefore, one needs to ensure the robustness of this method before applying it to multidimensional data that do not correspond to a correlated-traits model (e.g., bifactor or higher-order models) (Rios & Miranda, 2021).

Between-Person and Within-Person Subscore Reliability with Profile Analysis. Bulut et al. (2016) proposed a profile analysis-based method that can evaluate the reliability of both subscores and the total score. This method is based on the idea originally introduced by Davison et al. (2009) such that under the assumption of independence, a total variation in a test score profile partitions into two orthogonal components, between-person variation and within-person variation. The between-person variation means variations among examinees' subscore means (mean of the *K* subscores in the examinee's profile) that are also referred to as the *level* of the

subscore profile in the context of profile analysis. In contrast, the within-person variation means the variation between the subscores of each examinee which is also referred to as the *pattern* of the subscore profile. The pattern is operationalized as a vector of ipsatized subscores where each subscore is expressed as the deviation from its mean for a given examinee. Given that total variation in a test score profile equals the sum of between-person variation and within-person variation, the total profile reliability can be expressed as the weighted average of between-person and within-person reliabilities or, equivalently, as the weighted average of profile level and profile pattern reliabilities. Therefore, total profile reliability always rests in a range between within-person reliability and between-person reliability (Bulut et al., 2016).

According to the Bulut et al. (2016), the between-person and within-person reliabilities can be computed using the sample correlations between two parallel or tau equivalent test forms. If, however, no parallel test forms are available, one can create tau equivalent parcels by dividing each subtest into two item groups where two parcels should be matched as much as possible concerning length, difficulty, content, format, and other item features affecting response probabilities. Estimated profile reliability coefficients are interpreted similarly to other measures of internal consistency in the literature, such as coefficient alpha, KR-20, and marginal reliability in IRT.

Using simulated and real data, Bulut et al. (2016) showed that both between-person and within-person reliability are greatly affected by the intercorrelations among subscores. Specifically, as intercorrelations of subscores increase, the composite of subscores (i.e., the mean subscore for an examinee) gets more consistent and precise and thus the between-person reliability improves. On the contrary, as intercorrelations of subscores increase, the within-person variation diminishes, and hence the within-person reliability decreases. These results are expected given that conceptually the between-person reliability can be viewed as the total score reliability,

whereas the within-person reliability represents subscore orthogonality or distinctiveness to the rest of the test.

This approach is similar to Haberman's (2008) PRMSE and Wainer et al.'s (2001) augmentation method (described in detail in a later section) in that it involves the estimation of true score variances. However, the profile reliability approach estimates true score variances for both level and pattern from the sample correlations of parallel test forms or item parcels; thereby, it does not require the estimation of subtest score reliability coefficients for the estimation of profile reliability coefficients (Bulut et al., 2016). By contrast, Haberman's index estimates true score variances of subscores using observed variances of subscores, reliability of each subtest, and correlations among observed subscores. Moreover, the profile reliability approach utilizes the profile level (i.e., the mean subscore for each examinee) as a proxy for estimating observed and true within-person variations among subscores, whereas PRMSE uses the total test score as a proxy for estimating the true variances of observed subscores (Bulut et al., 2016).

One potential limitation in this approach is a lack of concrete decision rules for the level of reliability required for the level and pattern of the subscore profile. Given a substantial trade-off between within-person and between-person subscore reliability, Bulut et al. (2016) noted that the decision rule must be based on the purpose of a test being used. If test users are more interested in the relative strength and weaknesses of the individual, then acceptably high reliability of the profile pattern is desired. In contrast, if test users are more interested in the reliable distinction between individuals, then high reliability of the profile level is desired and, in that case, items may not provide any added values over and above the total score.

Criterion-Related Validity Evidence for Subscores. Whereas the aforementioned techniques examined the added value of subscores relying on information internal to the tests, Davison et al. (2015) utilized information external to the tests to examine the added value of

subscores. Their CTT-based method evaluates the contribution of subscores over total scores to the prediction of an external criterion, whereas Haberman's (2008) method evaluates the contribution of subscores over total scores to the prediction of true subscores. Davison et al.'s method is based on the criterion-related pattern analysis (CPA; Culpepper, 2009) procedure initially proposed by Davison and Davenport (2002). In this approach, the examinee's score on the criterion variable is expressed as the full model which is the standard linear multiple regression model with subscores as predictors:

$$Y_j = \sum_k b_k X_{jk} + a + \varepsilon_j, \tag{6}$$

where Y_j is the criterion variable for examinee j, b_k is the linear regression coefficient for subscore (predictor) k, X_{jk} is the observed subscore k for examinee j, a is the intercept value, and ε_j is the error term for examinee j. If subscores are not more predictive than the total score, the full model in Equation 6 will be reduced to a simple regression with a total score as the only predictor. This restricted model is a special case of the full model in which the regression weights would be constant across subscores:

$$Y_j = b \sum_k X_{jk} + a + \varepsilon_j. \tag{7}$$

The CPA can also be used for evaluating the extent to which the examinee's subscore profile pattern matches the criterion-related pattern (Davison et al., 2015; Davison & Davenport, 2002). The criterion-related pattern is defined as a vector of regression coefficients estimated in the full model expressed as the deviation from their mean. The magnitude of the match between the examinee's subscore pattern and the criterion-related pattern is indicated by a profile match statistic (Cov_j) that can be computed for each examinee as follows:

$$Cov_j = (1/K) \sum_{k=1}^{K} (b_k - \bar{b})(X_{jk} - \bar{X}_j), \tag{8}$$

where K is the number of subtests, and $\bar{b}$ is the mean regression coefficient, and $\bar{X}_j$ is the mean subscore for examinee j. A negative profile match statistic indicates the examinee's pattern is the reverse of the given criterion-related pattern.

Davison et al. (2015) have shown that the full model defined above can be reparameterized in terms of the total score and the profile match statistic (Cov_j) such that the predicted score on the criterion variable is expressed as the linear combination of the total score and the profile match statistic. If the restricted model holds, the total score is a sufficient predictor; thus, regardless of subscore pattern, every combination of subscores resulting in the same total score would lead to the same predicted score on the criterion variable. On the other hand, if the full model improves prediction, then both the total score and profile match statistic contribute to the prediction; thereby, examinees who have the same total score but different subscores are likely to have different predicted scores on the criterion variable. In this case, the examinee's predicted value on the criterion variable will depend on the extent to which the examinee's subscore profile pattern matches the criterion-related pattern (Davison et al., 2015).

According to Davison et al. (2015), the CPA procedure is advantageous in that it can estimate actual incremental validity evidence provided by a set of subscores and allows for a significance test of the statistical importance of this increment. Specifically, the incremental predictive validity evidence of subscores over and above the total score is evaluated using the F-test which compares R^2 between a full regression model (i.e., using subscores as predictors) and a restricted model (i.e., using the total score as the only predictor) relative to the difference in their degrees of freedom for error. Aside from the statistical significance, the effect sizes of incremental predictive validity evidence are indicated by the change in R^2 between the full and restricted regression models. Authors also noted that the CPA can be applied not only to the

continuous criterion variable but also to the discrete criterion variable using a nonlinear link function such as logit or probit (e.g., Chan, 2006; Davison et al., 2014; Huang, 2015).

Subscore Estimation Methods from CTT and IRT Frameworks

This section describes three most widely used CTT- or IRT-based subscore estimation methods which yielded more accurate subscores than other diagnostic methods (de la Torre & Patz, 2005; Dwyer et al., 2006; Fu & Qu, 2018). In what follows, their theoretical background and strengths and limitations are illustrated. This section concludes with a review of their comparative performances under various conditions.

Regression Approach

Theoretical Background of Regression Approach. The regression approach seeks to increase the precision of a subscore by borrowing information from out-of-scale items assuming that subscores are most likely to correlate at least moderately. To this end, this approach utilizes a linear regression to estimate classical true scores of subtests using observed scores as predictors. As in the typical regression approach, regression coefficients are estimated based on the least-squares method and classical true score theory (Fu & Qu, 2018). Subscore estimation methods under this approach include subscore augmentation (Wainer et al., 2001) and weighted averages (Haberman, 2008). These two methods differ in how they utilize observed scores as predictors.

Subscore augmentation (Wainer et al., 2001) is based on Kelly's (1927) regressed scores which are expressed by

$$\hat{T}_j = \rho_X X_j + (1 - \rho_X)\bar{T}, \tag{9}$$

where $\hat{T}_j$ is the estimated true score for examinee *j*, ρ_X is the reliability of the observed score *X*, X_j is the observed score for examinee *j*, and $\bar{T}$ is the mean true score. As $\bar{T} = \bar{X}$ under CTT assumptions, Equation 9 can be written as:

$$\hat{T}_j = \rho_X X_j + (1 - \rho_X)\bar{X}, \tag{10}$$

where $\bar{X}$ is the mean observed score. This method aims to improve the estimate of the true score using the group mean as the ancillary information; in essence, it removes the unreliable part of the observed score by regressing it to the group mean. According to Equation 10, $\hat{T}_j$ is more influenced by the observed score as the reliability of the observed score increases, whereas $\hat{T}_j$ is increasingly influenced by the group mean as the reliability of the observed score decreases.

Wainer et al. (2001) extended Kelly's (1927) method to multivariate cases in which a test comprises multiple subscales. To apply Wainer et al.'s method, Equation 10 is algebraically rearranged as follows:

$$\hat{T}_j = \bar{X} + \rho_X(X_j - \bar{X}) \tag{11}$$

and can be expressed in a multivariate case as:

$$\hat{\boldsymbol{T}}_{\boldsymbol{j}} = \bar{\boldsymbol{X}} + \boldsymbol{B}(\boldsymbol{X_j} - \bar{\boldsymbol{X}}), \tag{12}$$

where $\hat{\boldsymbol{T}}_{\boldsymbol{j}}$ is the vector of estimated true subscores for examinee *j*, $\boldsymbol{X_j}$ is the vector of observed subscale scores for examinee *j*, $\bar{\boldsymbol{X}}$ is the vector of subscale means for the sample of examinees, and $\boldsymbol{B}$ is a matrix of weights (i.e., regression coefficients). As such, Wainer et al.'s (2001) subscore augmentation predicts the true scores of each subtest using observed raw scores of all subtests.

The weight matrix $\boldsymbol{B}$ is the multivariate analog for the estimated reliability and can be estimated by:

$$\boldsymbol{B} = \sum_{\boldsymbol{T}} \sum_{\boldsymbol{X}}^{-\boldsymbol{1}}, \tag{13}$$

where $\sum_{\boldsymbol{T}}$ is the variance-covariance matrix of true subscale scores, and $\sum_{\boldsymbol{X}}^{-\boldsymbol{1}}$ is the inverse of the variance-covariance matrix of observed subscale scores. Equation 13 is based on the CTT definition of reliability as the proportion of true score variance relative to the observed score variance. Whereas $\sum_{\boldsymbol{X}}^{-\boldsymbol{1}}$ can be obtained directly from the observed subscale scores, $\sum_{\boldsymbol{T}}$ must be

estimated from $\sum_{\boldsymbol{X}}$. As true scores are assumed to be uncorrelated with errors in CTT, the off-diagonal elements of $\sum_{\boldsymbol{T}}$ will be equal to those of $\sum_{\boldsymbol{X}}$. On the other hand, the diagonal elements of $\sum_{\boldsymbol{T}}$ can be computed as the product of each diagonal element of $\sum_{\boldsymbol{X}}$ and the reliability of each subscale.

By contrast, Haberman's (2008) weighted average is computed by a weighted average of a subscore and the total score as follows:

$$\hat{T}_{jk} = \bar{X}_k + a\left(X_{jk} - \bar{X}_k\right) + b\left(X_j - \bar{X}\right), \tag{14}$$

where *a* and *b* are weights determined by standard deviations, reliabilities, and correlations of subscale scores and the total score. The weighted average is considered a special case of the augmented subscore; the weighted average imposes the same weight on all subscale scores except the target subscale, whereas the augmented subscore utilizes different weights on all subscale scores (Sinharay, 2010).

Strengths and Limitations of Regression Approach. The regression approach is relatively easy to implement and can be used in conjunction with various score types such as conventional summed scores, scale scores, and IRT scores (de la Torre et al., 2011). In addition, researchers have demonstrated that both subscore augmentation and weighted averages have improved reliability and added value of subscores although improvement varied from marginal (e.g., Papageorgiou & Choi, 2018) to substantial (e.g., Sinharay, 2010). According to Fu and Qu (2018), the degree of the improvement depends on the relevance and amount of the ancillary information which in turn is determined by the following factors: (1) reliability of an in-scale subtest; (2) reliability and test length of out-of-scale subtests; and (3) inter-subdomain correlations. In general, the improvement by the regression approach will be greater when the in-scale subtest has relatively low reliability and out-of-scale subtests are longer and hence more reliable and highly correlated to each other.

However, borrowing information from other subscores and the total score inevitably decreased the subscore distinctiveness (Feinberg & Wainer, 2014; Stone et al., 2010). Thus, as the reliability of raw subscores decreased, augmented subscores became more indistinguishable from the total score. Therefore, the regression approach would be appropriate when raw subscores have ample orthogonality (Sinharay et al., 2011). For these reasons, Sinharay et al. (2019) suggested guidelines regarding when to report statistically enhanced subscores: (a) simply report raw subscores if they are reliable and distinct from other subscores; (b) report neither raw nor statistically enhanced subscores if raw subscores are unreliable and not dimensionally distinct from other subscores; or (c) report statistically enhanced subscores if raw subscores are moderately reliable and are moderately correlated with other subscores.

In addition, as the regression approach involves a two-step process (i.e., initial subscores are first estimated and then augmented based on collateral information obtained from out-of-scale items), estimation errors involved in both steps are aggregated into the final subscore estimates (Chin, 2011). Besides, test users may not understand the rationale behind computing subscores based on the examinee's performance not only on the target subdomain but also on other subdomains (Fu & Qu, 2018). Sinharay et al. (2019) suggested that such an issue can be addressed by stressing the fact that the augmented subscore better reflects the examinee's ability in the target subdomain than does the observed subscale score.

Subscore Augmentation with Theta Estimates from an IRT Model

Theoretical Background of Subscore Augmentation with Theta Estimates from an IRT Model. In addition to CTT-based subscores, theta estimates of subtests calibrated from unidimensional IRT models have also been used as predictors in Wainer et al.'s (2001) subscore augmentation model (e.g., de la Torre et al., 2011; de la Torre & Patz, 2005; Edwards & Vevea, 2006; Fu & Qu, 2018; Wainer et al., 2001). Unlike the unidimensional IRT (UIRT) approach, the

augmented IRT (AIRT) method takes into consideration relationships between subdomains when estimating subscores. To apply this method, unidimensional IRT ability estimates must be first obtained using maximum likelihood (ML), maximum a posteriori (MAP), or expected a posteriori (EAP) method. Theta estimates obtained by maximum likelihood (ML) can be directly used as predictors in the regression model (Equation 12) in the same way as raw subscores. In contrast, if ability estimates obtained using Bayesian estimators such as MAP or EAP are applied, these estimates must be corrected before being used as predictors in the regression model due to their tendency to shrink toward the population mean (Fu & Qu, 2018).

Assuming that the population means of thetas on each subtest is zero and the standard errors of theta estimates are constant, the correction is made as follows:

$$\mathrm{MAP}^{*}(\theta_k) = \frac{\mathrm{MAP}(\theta_k)}{\rho_k}, \tag{15}$$

where $\mathrm{MAP}^{*}(\theta_k)$ is the corrected MAP estimate on subscale *k*, $\mathrm{MAP}(\theta_k)$ is the original theta estimate on subscale *k* calibrated from the unidimensional IRT model, and ρ_k is an estimate of the reliability of subscale *k*. Whereas the CTT-based subscore augmentation approach evaluates the reliability of subscales using coefficient alpha, AIRT utilizes marginal reliability which is estimated as follows:

$$\hat{\rho}_k = \frac{\sigma^2_{T_k}}{\sigma^2_{T_k} + \sigma^2_{e_k}}. \tag{16}$$

where $\hat{\rho}_k$ is the estimate of the marginal reliability for subscale *k*, $\sigma^2_{T_k}$ is the true theta variance for subscale *k*, and $\sigma^2_{e_k}$ is the error variance for subscale *k*. In turn, $\sigma^2_{T_k}$ and $\sigma^2_{e_k}$ are computed by:

$$\hat{\sigma}^2_{T_k} = \frac{\sum_{j=1}^{J}(\hat{\theta}_{jk} - \bar{\theta}_k)}{J-1} \text{ and } \hat{\sigma}^2_{e_k} = \frac{\sum_{j=1}^{J} SE^2_{\hat{\theta}_{jk}}}{J}, \tag{17}$$

where $\hat{\theta}_{jk}$ is the theta estimate on subscale k for examinee j, $\bar{\theta}_k$ is the mean theta estimate on subscale k, $SE^2_{\hat{\theta}_{jk}}$ is the squared standard error of each theta estimate on subscale k, and J is the number of examinees.

Strengths and Limitations of AIRT Method. In general, the strengths and limitations of the augmentation approach stated in the previous section apply to the AIRT as well. Accordingly, augmentation approaches using raw subscores and IRT theta estimates yielded comparable results. For example, Edwards and Vevea (2006) compared augmented subscores based on both EAP theta estimates and raw scores to non-augmented subscores (i.e., number correct for the subdomain). Data were simulated based on a three-parameter logistic (3PL) IRT model varying conditions on (a) the number of subdomains (2 and 4), (b) subscale lengths corresponding to four levels of coefficient alpha (5, 10, 20, and 40 items per subscale corresponding to coefficient alpha of .43, .59, .75 and .85), and (c) correlations among thetas (.3, .6, and .9). Evaluation criteria included (a) root mean square error (RMSE; square root of the average squared difference between true and estimated thetas), (b) reliability (square of the correlation between true and estimated thetas), (c) the percentage of simulees for whom estimated augmented scores are closer to true scores than are non-augmented scores, and (d) classification accuracy.

Results indicated that compared to non-augmented subscores, augmented subscores based on both raw subscores and theta ability estimates improved estimation and classification accuracy across all conditions. Consistent with other findings, the augmented subscores became more accurate as the length (i.e., reliability) of the in-scale subtest decreased and the length (i.e., reliability) and inter-correlations of the out-of-scale subtests increased. On the other hand, the number of subdomains had a negligible impact on results.

MIRT Approach

Theoretical Background of MIRT Approach. MIRT models assume that each subscale represents a distinct trait, and the final score reflects a mixture of related multidimensional traits. In the MIRT approach, item parameters are derived from the factor loadings of each item on each subscale as well as estimated covariances between subscores, and then person parameters are calculated based on these values (Reckase, 2009). As such, MIRT models allow for simultaneous estimation of all items considering the intercorrelation of subdomains, whereas UIRT models analyze items from each domain separately (Svetina et al., 2017). The MIRT approach to estimating subscores is similar to the augmented approach in that it utilizes out-of-scale item responses to improve the subscore reliability. However, instead of borrowing the information from subscores of other examinees, the MIRT approach borrows the information from the examinee's own scores on other subscales (Longabach & Peyton, 2018).

Psychometric Models of MIRT. Although researchers have proposed various MIRT models, such as the multidimensional Rasch model, the multidimensional two-parameter and three-parameter logistic models, the multidimensional graded response model, and the bi-factor IRT model, these models are often multidimensional extensions of the unidimensional IRT models. MIRT models can be grouped in several ways. For instance, MIRT models can be categorized based on their primary purposes. MIRT models can also be grouped depending on whether the weakness in one dimension can be compensated by the strength in another dimension. Additionally, MIRT models can be categorized based on whether each item measures a single latent trait or multiple latent traits. In what follows, these categorizations are further described.

Exploratory vs. Confirmatory MIRT Models. Exploratory MIRT models are employed when there is no a priori hypothesis for the test structure underlying item responses. In such situations, the expected number of dimensions is specified before estimating the item parameters

and abilities, whereas relationships between latent traits and items are derived from the data. Exploratory models are typically used for either determining the underlying dimensions of a test or checking the unidimensionality assumption. In contrast, confirmatory MIRT models are employed when there is a hypothesis about the test structure. Therefore, both the number of dimensions and relationships between latent traits and items are specified before estimating model parameters (Reckase, 2009). Exploratory and confirmatory MIRT models are analogous to exploratory and confirmatory factor analysis, respectively.

Compensatory vs. Non-compensatory MIRT Models. Compensatory models assume that weakness in one latent trait can be compensated by strength in another latent trait (Reckase, 1997; Yao & Boughton, 2007). Thus, examinees may use one of several alternative strategies for answering an item correctly rather than using one set of skills only. Given the additive nature, compensatory models define the probability of a correct response as a function of the weighted sum of a series of latent traits where weights are determined by discrimination parameters.

As an example of compensatory models, the multidimensional two-parameter logistic (M2PL) model (Reckase, 1985) takes the same general form as the unidimensional 2PL IRT model which links the latent ability and the probability of a correct item response through an item's discrimination and difficulty parameters as follows:

$$P\left(X_{ji}=1\middle|\theta_j,a_i,b_i\right)=\frac{\exp\left(Da_i(\theta_j-b_i)\right)}{1+\exp\left(Da_i(\theta_j-b_i)\right)}, \tag{18}$$

where θ_j is the latent trait for examinee j (j = 1, …, J), a_i and b_i are the item discrimination and item difficulty parameter for item i (i = 1, …, I), respectively, and D is a scaling constant (typically 1.7) used to make the logistic metric more closely aligned with the traditional normal ogive metric. Unlike the 2PL model, the M2PL model provides the discrimination parameter per each latent trait as well as the intercept term per item. This allows for a better understanding of

multiple latent traits associated with the item response. The M2PL expresses the probability that an examinee *j* with ability θ correctly answers a dichotomous item *i* as follows:

$$P\left(X_{ji}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)=\frac{\exp\left[D\left(\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i\right)\right]}{1+\exp\left[D\left(\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i\right)\right]}, \tag{19}$$

where $\boldsymbol{\theta_j}$ is a 1 × *K* vector of latent traits for examinee *j* $\left(\boldsymbol{\theta_j}=\ (\theta_{j1},\theta_{j2},\cdots,\theta_{jK})\right)$, $\boldsymbol{a_i}$ is a 1 × *K* vector of slope (i.e., item discrimination) parameters for item *i* $(\boldsymbol{a_i}=\ (a_{i1},a_{i2},\cdots,a_{iK}))$, d_i is an intercept term for item *i*, and *D* is a scaling constant as in Equation 18. It should be noted that the intercept parameter, d_i, is different from the item difficulty parameter in a UIRT model because this parameter is not a unique indicator of the item difficulty (Reckase, 2009). Under the simple structure, the traditional difficulty parameter (B_i) can be obtained from the following transformation using intercept and discrimination parameters:

$$B_i=\frac{-d_i}{\sqrt{\sum_{k=1}^{K}a_{ik}^2}}. \tag{20}$$

By contrast, non-compensatory models assume that the examinee must master all the skills required to answer an item correctly. The non-compensatory M2PL model for dichotomously scored items is expressed as follows:

$$P\left(X_{ji}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},\boldsymbol{b_i}\right)=\prod_{k=1}^{K}\frac{\exp\left[a_{ik}(\theta_{jk}-b_{ik})\right]}{1+\exp\left[a_{ik}(\theta_{jk}-b_{ik})\right]}, \tag{21}$$

where $\boldsymbol{b_i}$ is a 1 × K vector of difficulty parameters for item *i* $(\boldsymbol{b_i}=\ (b_{i1},b_{i2},\cdots,b_{iK}))$, and a_{ik}, b_{ik}, and θ_{jk} are discrimination, difficulty, and latent ability parameters for dimension *k* of item *i*, respectively. Unlike the compensatory model, this model includes separate difficulty parameters for each dimension and defines the probability of a correct response as a function of the multiplication of probabilities, each based on a different latent trait. Therefore, the weakness in one latent trait cannot be compensated by the strength in another latent trait; to obtain a high probability under non-compensatory models, high abilities for all dimensions are required

(Reckase, 1997, 2009). These models are more appropriate when the correct response to an item requires the successful execution of a series of steps in a specific order as in tests for mathematical abilities (Longabach & Peyton, 2018).

Although several applications of non-compensatory MIRT models exist in the cognitive assessment literature (Embretson, 1997; Junker & Sijtsma, 2001; Maris, 1995; Sympson, 1978; Whitely, 1980), non-compensatory models have not been widely used. This is primarily because non-compensatory models impose a computational burden by requiring the estimation of a separate difficulty parameter for each latent trait of the item (C. Wang & Nydick, 2015). Compensatory models, by contrast, have been popular in research and practice due to their relative simplicity in estimation and their theoretical similarities to factor analysis models (Sijtsma & Junker, 2006).

Between-Item vs. Within-Item MIRT Models. MIRT models can be also categorized based on whether each item is associated with a single latent trait (i.e., simple structure) or multiple latent traits (i.e., complex structure). The MIRT models only consisting of the simple structured-items are referred to as between-item MIRT models in which each subdomain is measured by a unique set of items. Between-item MIRT models are appropriate for estimating subscores from subtests measuring correlated latent traits (Desjardins & Bulut, 2018). By contrast, MIRT models containing complex-structured items are referred to as within-item MIRT models (Adams et al., 1997; W.-C. Wang et al., 2004) in which each subdomain is measured by a unique and shared set of items. Higher-order IRT or bi-factor IRT models are an example of within-item MIRT models in which each item measures one or more specific abilities plus a general ability.

Specification of Item-Trait Relationships in MIRT Models. As stated earlier, confirmatory MIRT models explicitly specify the relationship between the item and the latent

traits based on prior knowledge regarding the latent structure of the assessment (Finch, 2011). To clearly express the item–trait relationship in MIRT models, da Silva et al. (2019) incorporated the item–trait matrix, also known as the Q-matrix, into MIRT models. Although the item–trait relationship has been previously defined when implementing MIRT models, the use of Q-matrix has been largely limited to the DCM approach.

In the Q-matrix, q_{ik} is the element of row i and column k indicating whether item i measures dimension k or not: q_{ik} = 1 if the item i measures the dimension k, and q_{ik} = 0 otherwise. The q_{ik} can be incorporated into the latent linear predictor (i.e., $\eta_{ji} = \boldsymbol{a_i\theta_j}' + d_i = \sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i$) of the M2PL model in Equation 19 as follows:

$$P\left(X_{ji} = 1 \middle| \boldsymbol{\theta_j}, \boldsymbol{a_i}, d_i\right) = \frac{\exp(\boldsymbol{a_i\theta_j}' + d_i)}{1+\exp\ (\boldsymbol{a_i\theta_j}' + d_i)} = \frac{1}{1+\exp[-(\boldsymbol{a_i\theta_j}' + d_i)]}$$

$$= \frac{1}{1+\exp\left[-\left(\sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i\right)\right]} = \frac{1}{1+\exp\left[-\left(\sum_{k=1}^{K} q_{ik}a_{ik}\theta_{jk} + d_i\right)\right]}, \quad (22)$$

where a new latent linear predictor, η'_{ji}, is expressed as $\sum_{k=1}^{K} q_{ik}a_{ik}\theta_{jk} + d_i$. The resultant model, M2PL-Q, in Equation 22 is a generalization of other IRT models. For example, the M2PL-Q with the Q-vector consisting only of elements equal to 1 across all items is equivalent to the unidimensional 2PL model (da Silva et al., 2019).

According to da Silva et al. (2019), the Q-matrix not only accurately reflects the item and test design considerations but also provides some practical benefits. The constraints provided by the Q-matrix "decrease the number of discrimination parameters to be estimated for the model, simplify the estimation process, increase the accuracy of latent trait estimates, and considerably shorten the estimation time as compared to a fully unrestricted MIRT" (p. 683).

Strengths and Limitations of the MIRT Model. Many researchers (e.g., de la Torre, 2008b; de la Torre & Patz, 2005; Haberman & Sinharay, 2010; W.-C. Wang et al., 2004; Yao & Boughton, 2007) evaluated the quality of subscore estimates from the MIRT approach. In general,

they have found that ability estimates from MIRT are more accurate than raw subscores or theta estimates calibrated from UIRT which do not take into consideration relationships between subscales. As an example, de la Torre and Patz (2005) compared the accuracy and precision of multidimensional EAP estimates to those of unidimensional EAP estimates. Data were simulated based on the simple structure multidimensional three-parameter logistic (M3PL) model manipulating factors such as (a) the number of subtests (2 and 5), (b) the number of items (10, 30, and 50), and (c) correlation of thetas (0, .4, .7, and .9). They employed the hierarchical Bayesian approach with the Markov chain Monte Carlo (MCMC) technique for estimation. Results indicated that in the presence of non-zero correlations between abilities, the MIRT approach taking the correlation into account could noticeably improve the accuracy and precision of ability estimates when compared to its unidimensional counterpart. In particular, the greater improvement was observed when there were multiple short subtests and the underlying correlation was high.

As is the case with the augmentation approach, however, increased reliability in the MIRT approach comes with a price: it increases the reliability of subscores at the expense of increasing their correlation, thus, decreasing their distinctiveness (Bulut, 2013). Also, MIRT models have not demonstrated improved performance as to subscore estimation errors when compared to CTT or UIRT models (Luecht, 2003). In addition, the use of MIRT models in practice is somewhat limited due to heavy computational burden. The number of parameters that need to be estimated is much larger for MIRT models than for unidimensional IRT models. Moreover, combined with complex estimation techniques such as the Bayesian estimation using the MCMC technique, estimation of MIRT gets unstable for models with too many parameters and impossible or very slow to execute for large datasets due to the limitations of software and computer resources (Fu & Qu, 2018). Accordingly, MIRT models require lengthier tests and

larger sample sizes for accurate and precise estimates of model parameters and latent abilities on multiple domains (Haberman & Sinharay, 2010; Templin & Bradshaw, 2013).

Comparison of Subscore Estimation Methods Based on CTT or IRT

Researchers (e.g., de la Torre et al., 2011; de la Torre & Patz, 2005; Haberman & Sinharay, 2010) have indicated that in general, the augmentation and MIRT approaches yielded comparable subscore estimates that are more accurate than raw subscores and theta estimates from UIRT. For example, using simulated data, de la Torre et al. (2011) compared four subscore estimation methods including AIRT, MIRT, higher-order IRT, and the objective performance index (OPI; Yen, 1987) which employs Bayesian IRT estimation to obtain an estimated true score (i.e., estimated proportion correct) for items on a target subdomain given their overall test performance. Manipulated factors included (a) the number of subtests (2 or 5), (b) subtest length (10, 20, or 30 items per subtest), and (c) correlation between thetas (0, .4, .7, or .9). The recovery of ability parameters was compared between four methods using bias, RMSE, correlations, and estimated proportion correct. Across all conditions, MIRT and AIRT were found to yield comparable results although the MIRT approach performed better for extreme thetas. The comparable performance of MIRT and AIRT methods is attributable to the fact that both methods incorporate the inter-subdomain correlations into the subscore estimation rather than assuming zero or perfect correlations between subscores as is the case with UIRT (Fu & Qu, 2018).

Although the MIRT approach can produce accurate theta estimates under adequate conditions for stable calibration (e.g., a large number of items and sample sizes), its heavy data demand and computational burden may make the augmentation approach preferable over the MIRT approach in practice. On the contrary, MIRT models are theoretically more appealing than the augmentation approach given their capability of simultaneously estimating all subscores without going through two-step procedures, which leads to a further reduction in the bias and

standard error of the estimates. Moreover, although the augmentation approach can yield estimates comparable to those of the MIRT model in the case of simple structure, the MIRT approach may be extended more easily to accommodate the complex structure and a richer set of collateral information (de la Torre & Patz, 2005).

Diagnostic Classification Models

Given the growing demand for diagnostic information, cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014; Nichols et al., 1995) have attracted increasing attention in the past decades in the field of educational measurement. CDAs are typically designed to measure a set of finer-grained skills which are usually referred to as attributes. The primary purpose of CDAs is to provide examinees with attribute profiles from which their strengths and weaknesses can be inferred. To this end, CDAs usually depend on diagnostic classification models (DCMs). To provide a comprehensive review of DCMs, this section consists of the following five subsections: (a) properties of DCMs distinct from other diagnostic methods; (b) attribute specification; (c) psychometric models of DCMs and model selection procedure; (d) quality of DCM estimates; and (e) advantages and limitations of DCMs.

Properties of DCMs Distinct from Other Diagnostic Methods

DCMs are distinct from CTT, IRT, or MIRT in several aspects. First, DCMs assume categorical latent variables and model the probability of correctly answering an item as a function of an attribute mastery pattern (Henson & Douglas, 2005). By contrast, CTT, IRT, or MIRT assume continuous latent variables and model the probability of correctly answering an item as a function of the examinee's rank order on a single or multiple continuous latent abilities.

Second, whereas CTT, IRT, or MIRT provide a total score or a profile of subscores, DCMs provide attribute mastery profile indicating the presence (1) or absence (0) of multiple fine-grained skills and classification of examinees into discrete latent classes based on that profile

(de la Torre & Douglas, 2008; Rupp et al., 2010). The presence (i.e., mastery) or absence (i.e., non-mastery) of skills are determined by the posterior probability of mastery (PPM; also known as the marginal probability of attribute mastery) for each skill. Specifically, examinees with the PPM equal to or greater than .5 are classified as masters on that attribute and otherwise classified as non-masters (Rupp et al., 2010).

Third, as in CFA models, item–trait relationships are explicitly defined a priori through the Q-matrix in DCMs, which is especially advantageous in the case of complex data structures. Based on these unique properties, Rupp and Templin (2008b) defined DCMs as "… probabilistic, confirmatory multidimensional latent variable models with a simple or complex loading structure" (p. 226).

Fourth, DCMs assume equal difficulty across all items within a test, whereas CTT, IRT, or MIRT typically assumes nonzero variability for item difficulty parameters (de la Torre & Lee, 2010; Roussos et al., 2005). As DCMs assume discrete latent traits, thresholds between categories are assumed to be the same for all items (Chin, 2011). It is noteworthy that the equal difficulty assumption is advantageous for tests designed to provide criterion-referenced rather than norm-referenced interpretation, which is the case for DCMs. Norm-referenced tests typically require a wide range of item difficulty to determine each examinee's relative standing on a continuous scale, whereas diagnostic tests require that item difficulty should center at the cut-scores with a smaller variability to classify examinees (Jang, 2009; X. Wang et al., 2019).

Attribute Specification

Attribute specification reflects the substantive theory on which the diagnostic assessment in use was constructed. It is a critical component in any DCMs because the validity of the inferences made about the examinee's performance depends on the accuracy of the attribute specification (Rupp et al., 2010). In what follows, attribute specification of DCMs is illustrated

focusing on (a) definitional grain size and the number of attributes, (b) construction and validation of Q-matrix, (c) attribute hierarchies, and (d) complex versus simple item structure.

Definitional Grain Size and the Number of Attributes. The definitional grain size of the attributes refers to the degree of definitional specificity of attributes and in general increases as the scope and cognitive complexity of the task being analyzed increases (Rupp et al., 2010). The grain size of the attributes should be considered from both theoretical and practical perspectives (Jang, 2009; Rupp et al., 2010). In the theoretical perspective, attributes must be as specific as possible to adequately represent the cognitive theory underlying item response processes so that diagnostic feedbacks can better inform instructional and remedial effort. When the scope of an attribute is underrepresented, examinees may be misclassified (Jurich & Bradshaw, 2014). In the same vein, Jang (2009) noted that attributes as few as three to five are insufficient to cover the content and underlying construct.

From a practical standpoint, however, as attributes are more finely defined, their estimation and interpretation become more challenging (Embretson & Yang, 2013; G. Xu & Zhang, 2016). That is, more finely defined attributes increase the number of attributes and model parameters to be estimated, which in turn increases sample sizes and the number of items required for accurate and stable estimation of model parameters. In particular, when coupled with complex models (i.e., general or saturated DCMs) which typically involve more parameters to be estimated, a large number of attributes will cause issues like poor model identification, longer computational time, and resource constraints (Ravand & Baghaei, 2019; Templin & Bradshaw, 2013). Researchers have also shown that as the number of attributes measured by the whole test and each item increased, the bias and standard errors of parameter estimates increased as well (DiBello et al., 2007; Skaggs et al., 2016).

A larger number of attributes also decreases the number of items measuring each attribute for a given test, which may in turn lead to lower attribute reliability. As to the number of items measuring each attribute, Hartz (2002) recommended at least three items for any attribute, whereas Jang (2009) suggested at least five items per attribute. It should be noted that not only the number of items but also the quality of items impacts the psychometric quality of attribute estimates. For example, if an attribute is measured by only a few items with varying levels of discrimination, the examinee's performance on the highly discriminating item would largely determine the classification: examinees who correctly answer that item are likely to be classified as masters of the attribute whereas those who miss that item are likely to be classified as non-masters (Ravand & Baghaei, 2019).

Considering these constraints, Ravand and Baghaei (2019) suggested that more fine-grained attributes should only be used along with simpler DCMs and a larger number of items and examinees. Besides, as a rule of thumb, de la Torre and Minchen (2014) recommended maximum10 attributes for a test. However, if the sample size is not large enough, attributes as many as 10 are likely to lead to sparse latent classes with most examinees assigned to one of two flat profiles (i.e., all mastered or none mastered) (Lee & Sawaki, 2009; Li, 2011; Ravand, 2015). This is because the number of latent classes to be estimated exponentially increases with the number of attributes (i.e., 2^K latent classes for K attributes). For these reasons, most DCM researchers have specified three to five attributes (R. Liu et al., 2018).

Construction and Verification of the Q-Matrix. As stated earlier, attribute specification in DCMs is encoded in a Q-matrix which has an element q_{ik} for I items and K attributes indicating whether mastery of attribute k is required by item i (i.e., $q_{ik} = 1$) or not (i.e., $q_{ik} = 0$) (Henson & Douglas, 2005). The Q-matrix is equivalent to the factor pattern matrix that assigns the loadings in CFA (Rupp et al., 2010). The Q-matrix is usually developed by domain

experts and assumed to be correct in the subsequent CDM analyses. As such, Q-matrix construction involves some degree of subjective judgment that potentially leads to misspecifications. These misspecifications can affect the quality of item parameter estimates, and ultimately the examinee classification accuracy (de la Torre, 2008; Rupp & Templin, 2008a). For this reason, expert-based Q-matrices need to be empirically verified to ensure accuracy in the inferences derived from CDMs.

To date, several empirical methods of Q-matrix validation have been proposed. According to Ravand and Baghaei (2019), these methods are categories into two types: (a) completely data-driven methods (e.g., Barnes, 2010; Chen et al., 2015, 2017; J. Liu et al., 2012) which derive underlying attributes from item responses and (b) methods used in conjunction with the expert-defined provisional Q-matrices to detect misspecifications in the Q-matrix (e.g., de la Torre, 2008a; de la Torre & Chiu, 2016; DeCarlo, 2012; Templin & Henson, 2006) or to assist derivation of the Q-matrix from item responses (e.g., Close, 2012; Desmarais & Naceur, 2013). For example, de la Torre and Chiu (2016) proposed a method based on the G-DINA discrimination index (GDI) that can be used with a wide class of CDMs. Given a provisional Q-matrix and item responses, the procedure identifies the correct q-vector for each item. A q-vector is deemed correct if it is the simplest q-vector and the proportion of variance accounted for (PVAF) by the q-vector is high relative to the maximum GDI of the item.

In addition to these methods, fit indices can also guide the selection of the best performing Q-matrix among the competing ones (Ravand & Baghaei, 2019). However, as noted by Lei and Li (2016), the true Q-matrix may not be among the matrices studied. Therefore, Lei and Li suggested that the interpretability of the attributes must be considered along with the fit indices to select the optimal Q-matrix. The selected Q-matrix should be also replicated with other

samples to ensure that the selected Q-matrix closely resembles the true one (Ravand & Baghaei, 2019).

Attribute Hierarchies. Attribute hierarchies are specifications of the attribute dependencies in the target population (Rupp et al., 2010). When attribute hierarchies are present, skills are sequentially mastered (e.g., learning new skills builds upon pre-requisite skills). Structural relationships among those skills can be modeled by the hierarchical DCMs (HDCMs) which impose a one-factor correlation structure estimating attribute association through the use of a higher-order trait as follows (Templin et al., 2008):

$$\rho_{kk'} = \lambda_k \lambda_{k'}, \tag{23}$$

where $\rho_{kk'}$ is the correlation between any pair of attributes, k and k', and λ_k and $\lambda_{k'}$ represents the loading of attribute k and k' onto the higher-order trait, respectively. Therefore, the higher-order approach requires only *K* parameters for modeling the correlations of the attributes, whereas the general approach assuming an unconstrained correlation matrix requires $K(K-1)/2$ parameters (Templin et al., 2008).

As such, correct specification of attribute hierarchies not only designates which attribute profiles should be observed and which attribute profiles would be illogical in a sample but also greatly reduces the number of parameters to be estimated and hence model complexity for identification purposes (de la Torre & Douglas, 2004; Templin & Bradshaw, 2014). Templin et al. (2008) have shown that hierarchical modeling of the joint distribution of the latent skills is robust to the multivariate normality assumption of the attributes and comparable to the general approach for classification and item parameter estimation accuracy. The higher-order model is also considered more parsimonious than the multivariate normal model in that it is based on the concept that the acquisition of specific knowledge is affected by general abilities (Close, 2012).

Complex versus Simple Item Structure. The data structure has a substantial impact on the classification accuracy in DCMs. The simple structure can yield higher estimation and classification accuracy when compared to the complex structure because it offers less confounded information about latent traits (da Silva et al., 2019). For instance, Madison and Bradshaw (2015) have found that for the same number of items, the classification accuracy increases in the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) as more attributes are measured based on simple structure. Moreover, researchers have shown that each attribute must be measured by at least one simple-structured item to identify all possible examinees' attribute patterns (Chiu et al., 2009; Chiu & Seo, 2009; DeCarlo, 2012). As the number of attributes required per item increases, DCMs are less likely to be identifiable (DiBello et al., 2007). Consequently, Rupp et al. (2010) indicated that up to two or three attributes are typically measured by each item in most operational testing situations although the whole test may measure more attributes.

Psychometric Models of DCMs and Model Selection Procedure

To date, various DCMs have been proposed, and most of them comprise two major components: Q-matrix and condensation rule (Maris, 1999). The condensation rule defines how attributes interact to yield manifest item responses. Based on specific condensation rules DCMs are formulated to accommodate, DCMs can be categorized into three groups: conjunctive, disjunctive, and compensatory models (Ma & de la Torre, 2020). More recently, general or saturated DCMs subsuming a variety of specific DCMs have also been proposed. In the subsequent section, the notation used for the remainder of this section is first presented followed by the description of general and specific DCMs. This section concludes with issues to be considered when selecting the diagnostic model for the data in use.

Notation. Let X_{ji} denote the binary item response of examinee *j* to item *i* (1 = correct, 0 = incorrect). In addition, the vector $\boldsymbol{\alpha}$ has an entry α_{jk} which indicates whether examinee *j* has mastered attribute *k* or not (1 = mastery, 0 = non-mastery). The vector $\boldsymbol{\alpha}$ can take 2^K distinct values indexing the 2^K latent classes in a corresponding DCM. On the other hand, the vector $\boldsymbol{Q}$ has an entry q_{ik} which denotes whether attribute *k* is required by item *i* (1 = required, 0 = not required). Then, the probability of examinee *j* with attribute profile $\boldsymbol{\alpha}_c$ answering item *i* correctly is denoted by $P(X_{ji} = 1|\boldsymbol{\alpha}_c) = P_i(\boldsymbol{\alpha}_c)$.

When not all attributes are required for item *i*, 2^K latent classes can be collapsed into $2^{K_i^*}$ latent groups, where $K_i^* = \sum_{k=1}^{K} q_{ik}$ indicates the number of attributes required for item *i*. To simplify the notation, let $\boldsymbol{\alpha}_{li}^*$ denote the reduced attribute vector consisting of the columns of the required attributes for item *i*, where $l = 1, \ldots, 2^{K_i^*}$. In this case, latent class *c* is collapsed into latent group *l*; thus, $P_i(\boldsymbol{\alpha}_c) = P(\boldsymbol{\alpha}_{li}^*)$ (Ma & de la Torre, 2020).

General or Saturated DCMs. To date, general DCMs which can accommodate a wider range of underlying processes have been proposed including Bayesian inference networks (BINs; Almond et al., 2007), general diagnostic model (GDM; von Davier, 2008), log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), and generalized DINA (G-DINA; de la Torre, 2011). Saturated DCMs typically include all possible effects of attributes (i.e., main effects and all possible interaction effects).

For example, the item response function (IRF) of the G-DINA model is expressed as:

$$g[P(\boldsymbol{\alpha}_{li}^*)] = \lambda_{i0} + \sum_{k=1}^{K_i^*} \lambda_{ik}\alpha_{lk} + \sum_{k'=k+1}^{K_i^*} \sum_{k=1}^{K_i^*-1} \lambda_{ikk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{i12\ldots K_i^*} \prod_{k=1}^{K_i^*} \alpha_{lk}, \quad (24)$$

where $g[P(\boldsymbol{\alpha}_{li}^*)]$ represents $P(\boldsymbol{\alpha}_{li}^*)$, $log[P(\boldsymbol{\alpha}_{li}^*)]$, and $logit[P(\boldsymbol{\alpha}_{li}^*)]$ when the identity, log, and logit links are used, respectively. λ_{i0} is the intercept for item *i*, λ_{ik} is the main effect of attribute *k* for item *i*, $\lambda_{ikk'}$ is the two-way interactional effect of attribute k and k' for item *i*, and

$\lambda_{i12\ldots K_i^*}$ is the higher-order interaction among $\alpha_1, \alpha_2, \ldots, \alpha_{lK_i^*}$ for item i. de la Torre (2011) has shown that several widely used DCMs can be obtained by imposing appropriate constraints on the G-DINA model. In what follows, those specific models are described.

Conjunctive DCMs. Like non-compensatory MIRT models, conjunctive DCMs assume that a deficit in one attribute cannot be compensated for by mastery of other attributes; thus, non-mastery of even a single required attribute can dramatically decrease the probability of a correct response (e.g., math test, verbal tasks such as verbal classification and synonym items, and Raven's Progressive Matrices) (Henson et al., 2009). Common conjunctive models include deterministic input, noisy "and" gate (DINA; Junker & Sijtsma, 2001) model, noisy input, deterministic "and" gate (NIDA; Junker & Sijtsma, 2001) model, and reduced versions of the reparameterized unified model (rRUM; Hartz et al., 2002).

As an example, the DINA model can be obtained by constraining all terms in the identity link G-DINA model to zero except the intercept and highest order interaction effects (Ma & de la Torre, 2020). Its IRF is hence given by:

$$P(\boldsymbol{\alpha}_{li}^*) = \lambda_{i0} + \lambda_{i12\ldots K_i^*} \prod_{k=1}^{K_i^*} \alpha_{lk}. \tag{25}$$

This results in two parameters – guessing ($g_i = \lambda_{i0}$) and slipping ($s_i = 1 - [\lambda_{i0} + \lambda_{i1\ \ldots K_i^*}]$) parameters – for each item regardless of the number of attributes required. The slipping parameter functions as an upper asymptote given that an examinee who mastered all the attributes required by an item may slip and lower the probability of a correct response by s_i. By contrast, similar to the c_i parameter of the 3PL, the guessing parameter functions as a lower asymptote given that an examinee who has not mastered any attributes required by an item may guess and increase the probability of a correct response by g_i.

Disjunctive DCMs. Whereas conjunctive models require the mastery of all attributes for the item, disjunctive models assign the highest probability of correct answers to examinees

mastering at least one of the required attributes (Chiu et al., 2009). That is, under disjunctive DCMs, the probability of a correct response is equal for an examinee possessing one required attribute on item *i* and an examinee possessing all required attributes on item *i*. Therefore, disjunctive models are suitable when there is more than one strategy that can be used to solve the item and each strategy may require a different skill (Henson et al., 2009).

As an example, Templin and Henson (2006) proposed the deterministic input, noisy output "or" gate (DINO) model. The DINO model can be obtained by constraining absolute values of all the main and interaction effect parameters in the G-DINA model to different signs by the order of interactions (R. Liu et al., 2018). Its IRF is expressed as:

$$P(\boldsymbol{\alpha}_{li}^{*}) = \lambda_{i0} + \lambda_{ik}\alpha_{lk}, \tag{26}$$

where $\lambda_{ik} = (-1)\lambda_{ik'k''} = \cdots = (-1)^{K_i^{*}+1}\lambda_{i12\ldots K_i^{*}}$ for $k = 1, \ldots, K_i^{*}$, $k' = 1, \ldots, K_i^{*} - 1$, and $k'' > k', \ldots, K_i^{*}$. As in the DINA model, this results in two parameters for each item regardless of the number of attributes required.

Compensatory DCMs. Compensatory DCMs assume that a deficit in one attribute can be compensated for by mastery of other attributes. Common compensatory DCMs include a special case of the GDM referred to as the compensatory reparameterized unified model (C-RUM; Hartz, 2002) and additive CDM (ACDM; de la Torre, 2011). In compensatory DCMs, latent variables operate in an additive fashion such that the link function is monotonically related to the probability of a correct response (Chiu et al., 2009). Therefore, compensatory models allow for more equivalent classes of response probabilities per item than does a conjunctive model, but they are not as specific as the conjunctive model in the way they model the response process (Chiu et al., 2009).

As an example, the ACDM can be obtained by constraining all interaction terms in the identity-link G-DINA model to zero. Its IRF is expressed as:

$$P(\boldsymbol{\alpha}_{li}^{*}) = \lambda_{i0} + \sum_{k=1}^{K_i^{*}} \lambda_{ik}\alpha_{lk}. \quad (27)$$

The ACDM assumes that each required attribute contributes to the success probability uniquely and independently. This model has $K_i^{*} + 1$ parameters for each item.

Model Selection. Selecting the right model is of critical importance because it affects the classification of examinees (Lee & Sawaki, 2009), which is the primary purpose of the DCMs. Concerning the general versus specific DCMs, general DCMs are advantageous in that they require fewer assumptions, so they are more likely to fit the data. As noted earlier, however, these more flexible frameworks are also more complex with more data demand and thus more likely to cause issues in identifiability, item parameter recovery, and classification accuracy (de la Torre, 2011; Kunina-Habenicht et al., 2012). In addition, fitting a model that is too complex relative to the data may result in overfitting; such a model would not reduce the complexity of the data structure enough to justify its use as an effective explanatory means (Rupp et al., 2010). By contrast, specific DCMs can be used with less data demand and provide a more straightforward interpretation (Huebner, 2010). However, fitting a model that is structurally simpler than the given data may result in a poor model fit (Rupp et al., 2010).

According to Ravand and Baghaei (2019), one can use one of the following three approaches to the model selection in DCMs: (a) applying a single specific DCM across all items of a test; (b) applying several specific models and selecting the best-fitting model; and (c) applying a general model and allowing each item to have its own best-fitting model, hence resulting in multiple specific models within a single test. Ravand and Baghaei recommended the third option if the sample size is large enough (> 5,000) and the second option otherwise.

When the third option is adopted, several techniques can be used to evaluate the item-level model fit. For example, the procedure proposed by Henson et al. (2009) can help select the best specific model for the item based on visual inspection. In addition, the Wald test (de la Torre,

2011; de la Torre & Lee, 2013) can help determine the best DCM for an item by statistically comparing the fit of the G-DINA model to that of several specific DCMs such as DINA, DINO, and ACDM under the null hypothesis that the specific and the general model fit equally well. Ma et al. (2016) found that choosing the best item-level DCM based on the Wald test resulted in higher classification accuracy when compared to fitting a single general or specific DCM to all items.

Psychometric Quality of Attribute Estimates

DCMs primarily produce two different kinds of examinee estimates which can be considered a type of subscores: estimates of binary attribute mastery (or a profile of binary attribute mastery) and PPM ranging from 0 to 1. As with any other subscores, reliability and validity evidence of these attribute estimates should be established before they are reported to stakeholders.

Classification Accuracy and Consistency. The reliability of most DCM estimates cannot be evaluated based on the standard definition of reliability assuming a continuous, unidimensional latent trait (Jang, 2009) although different approaches commonly employ the overall concept of partitioning the total variance into true and error (da Silva et al., 2019). There has been a paucity of studies where the reliability of DCM estimates was reported until reliability-like measures for attribute estimates have been recently proposed by several researchers (e.g., Cui et al., 2012; Gierl et al., 2009; Johnson & Sinharay, 2018, 2020; Templin & Bradshaw, 2013; W. Wang et al., 2015). The reliability of discrete attribute estimates is defined by classification accuracy or classification consistency. Classification accuracy is defined as the probability that the estimated classification corresponds to the true classification, whereas classification consistency is defined as the probability that two parallel forms of the assessment result in the same estimated classification (Sinharay & Johnson, 2019). In what follows, the

notation by Sinharay and Johnson (2019) used for the remainder of this section is first presented. Then, measures for attribute-level classification accuracy and consistency are described.

Notation. Suppose an assessment includes *I* items and measures *K* binary attributes. Let the vector $\boldsymbol{A} = (A_1, A_2, \ldots, A_K)'$ denote the true attribute mastery pattern for the examinee, where each element, A_k represents the binary latent variable indicating whether or not a randomly chosen examinee truly has mastered the attribute *k*. Also, let Ω denote the set of all possible values of **A** comprising 2^K attribute patterns and let vector $\boldsymbol{\alpha} = (a_1, a_2, \ldots, a_K)'$ denote the estimated attribute mastery pattern with a_k being a realization of A_k. The random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_i, \ldots, X_I)^T$ represents the item scores of a randomly chosen examinee on the test with the number of items equal to *I*. Then, $\hat{a}_k(x)$ indicates a binary estimate of a_k based on the set of observed item responses $\boldsymbol{x} = (x_1, x_2, \ldots, x_I)^T$, and $\hat{\boldsymbol{a}}(\boldsymbol{x}) = (\hat{a}_1(\boldsymbol{x}), \hat{a}_2(\boldsymbol{x}), \ldots, \hat{a}_K(\boldsymbol{x}))'$ indicates the vector of all attribute estimates for an examinee.

Classification Accuracy for Each Attribute. W. Wang et al. (2015) proposed agreement measures at the single attribute level based on Cui et al.'s (2012) procedure measuring agreement at the attribute pattern level. They defined the attribute-level classification accuracy as follows:

$$P_{CA_k} = P(\hat{a}_k(\boldsymbol{X}) = a_k). \tag{28}$$

This measure can be computed using a hypothetical 2×2 table (Table 2.1) where proportions of the examinee population are cross-classified by their true attribute indicator A_k and their estimated attribute $\hat{a}_k(x)$. If the true proportion, p_{mn}, is equal to

$$p_{mn} = P(A_k = m, \hat{a}_k(\boldsymbol{X}) = n), \tag{29}$$

then the attribute-level accuracy is computed as

$$P_{CA_k} = p_{00} + p_{11}. \tag{30}$$

Assuming that item parameters and Q-matrix are known, the cell proportion p_{mn} can be computed by averaging the mean posterior probability (i.e., $P(A_k = m|X = x)$) over all possible

item response patterns, 2^K. Therefore, the computation of this measure gets too intense with a large number of items in a test. To ease the computational burden, p_{mn} can be obtained by summing only over the response patterns observed in a random sample of examinees (Johnson & Sinharay, 2018; W. Wang et al., 2015).

Table 2.1

The Contingency Table for True Versus Estimated Mastery Category Proportions

	Estimated Attribute		
True Attribute	0	1	Total
0	p_{00}	p_{01}	p_{0+}
1	p_{10}	p_{11}	p_{1+}
Total	p_{+0}	p_{+1}	1

Classification Consistency for Each Attribute. W. Wang et al. (2015) defined the classification consistency at the attribute level as follows:

$$P_{CC_k} = P(\hat{a}_k(\boldsymbol{X}_1) = \hat{a}_k(\boldsymbol{X}_2)), \tag{31}$$

which indicates the probability that a randomly selected examinee would have the same attribute estimate on two parallel forms of a test (Sinharay & Johnson, 2019). Like the classification accuracy, classification consistency for each attribute evaluates agreement in a 2×2 table (Table 2.2) where proportions of the examinee population are cross-classified by their estimates on two parallel forms of a test. If the true correlation, r_{mn}, is equal to

$$r_{mn} = P(\hat{a}_k(\boldsymbol{X}_1) = m,\ \hat{a}_k(\boldsymbol{X}_2) = n), \tag{32}$$

then, the attribute-level classification consistency is computed as:

$$P_{CC_k} = r_{00} + r_{11}. \tag{33}$$

Table 2.2

The Contingency Table Classifying Individuals on Two Parallel Forms

Estimate from Parallel Form 1	Estimate from Parallel Form 2		Total
	0	1	
0	r_{00}	r_{01}	r_{0+}
1	r_{10}	r_{11}	r_{1+}
Total	r_{+0}	r_{+1}	1

Assuming that item parameters are known and item responses are independent conditional on the true latent attribute vector $\boldsymbol{a}$, these proportions can be obtained by applying the following (Sinharay & Johnson, 2019):

$$r_{mn} = P(\hat{a}_k(\boldsymbol{X}_1) = m, \hat{a}_k(\boldsymbol{X}_2) = n)$$

$$= \sum_{a} P(\hat{a}_k(\boldsymbol{X}_1) = m, \hat{a}_k(\boldsymbol{X}_2) = n|\boldsymbol{A} = \boldsymbol{a})\, P(\boldsymbol{A} = \boldsymbol{a})$$

$$= \sum_{a} P(\hat{a}_k(\boldsymbol{X}_1) = m|\boldsymbol{A} = \boldsymbol{a}) P(\hat{a}_k(\boldsymbol{X}_2) = n|\boldsymbol{A} = \boldsymbol{a})\, P(\boldsymbol{A} = \boldsymbol{a}).$$

Alternative Measures of Classification Accuracy and Consistency. Given that the proportion of agreement in 2 × 2 tables tends to overestimate classification accuracy even when the true and estimated classification are independent, Johnson and Sinharay (2018) suggested the following popular indices as alternative measures of classification accuracy:

- Youden's (1950) statistic compares the probability that a DCM correctly classifies an examinee who mastered the skill (true positive) to the probability that the DCM incorrectly classifies an examinee who did not master the skill (false positive) (Sinharay & Johnson, 2019). It can be easily computed using the cell proportions presented in Table 2.1 and ranges from -1 to +1.

- Lambda of L. A. Goodman and Kruskal (1954) is a measure of proportional reduction in error in cross-tabulation analysis. In applications of DCMs, it adjusts for a baseline case in which the DCM would always choose the modal category (Sinharay & Johnson, 2019). This statistic ranges from 0 (i.e., the posterior mode never differs from the prior modal classification) to 1 (i.e., there is perfect classification).
- Cohen's (1960) kappa (κ) calculates the difference between the observed and expected agreements under the independence assumption and normalizes that difference as follows (Sinharay & Johnson, 2019):

$$\kappa = \frac{p_{11}+p_{00}-p_{1+}p_{+1}-p_{0+}p_{+0}}{1-p_{1+}p_{+1}-p_{0+}p_{+0}}. \quad (34)$$

  This statistic ranges from 0 (i.e., independence between true and estimated classifications) and 1 (i.e., perfect classification).
- Tetrachoric correlation is an index of the association between two binary variables. It is computed by the correlation between two correlated normal random variables resulting in quadrant probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$ (Sinharay & Johnson, 2019). Based on the traditional concept of reliability, the squared tetrachoric correlation between the true and estimated attribute estimates represents the classification accuracy, whereas (non-squared) tetrachoric correlation between attribute estimates of parallel forms represents the classification consistency.
- Sensitivity refers to a true positive rate (i.e., the proportion of examinees truly belonging to the mastery category who are correctly classified to the mastery category). In contrast, specificity refers to a true negative rate (i.e., the proportion of examinees truly belonging to the non-mastery category who are correctly classified to the non-mastery category). Using the proportional entries in Table 2.1, they are computed as follows:

$$\text{Sens} = P(\hat{a}_k(\boldsymbol{X}) = 1 | A_k = 1) = \frac{p_{11}}{p_{1+}}, \text{ and} \tag{35}$$

$$\text{Spec} = P(\hat{a}_k(\boldsymbol{X}) = 0 | A_k = 0) = \frac{p_{00}}{p_{0+}}. \tag{36}$$

All of these agreement measures can be easily adjusted to the case of DCMs. They can indicate not only the classification accuracy but also the classification consistency when estimates of r_{mn} instead of p_{mn} are used (Sinharay & Johnson, 2019).

Validity Evidence of DCM Estimates. Given that DCMs are relatively new techniques, at present, there is a lack of studies where authors reported the validity evidence of DCM estimates with a few exceptions (e.g., de la Torre et al., 2018; Garcia et al., 2014; Jang, 2009; Jang et al., 2013a, 2015; Kunina-Habenicht et al., 2017; Kunina et al., 2008). For example, using operational data, Kunina et al. (2008) have compared the skill profiles from a multidimensional Rasch model to their counterparts from DCMs and examined their criterion-related validity using school grades. In another instance, Jang (2009) analyzed the non-diagnostic reading comprehension test data using the Fusion model, a type of compensatory DCM, and evaluated the validity of resultant diagnostic feedback based on both quantitative (i.e., the association between PPM and students' self-assessment on skill proficiency) and qualitative evidence (i.e., students' perception of their attribute profiles). Similarly, Jang et al. (2013) evaluated the correlation between the examinee's exposure to English and their DCM estimates for the English Language Learner reading skills. More recently, Kunina-Habenicht et al. (2017) conducted an empirical study to evaluate the incremental validity evidence of DCMs over and above the simpler unidimensional IRT approach where DCMs were applied to the assessment designed to provide diagnostic feedback at a finer grain size. In these studies, validity evidence at least partially supported diagnostic information provided by DCMs.

It is noteworthy that as in these studies, most researchers urging more validity evidence for DCM applications underscored the importance of criterion-related validity evidence (Sessoms

& Henson, 2018) which concerns the extent to which estimated skill masteries are correlated with other relevant variables in theoretically expected ways. This is in part because more formal testing of subscore added value such as Haberman's (2008) PRMSE is not directly applicable to DCMs given that DCMs are based on the assumptions distinct from those of CTT or IRT. Moreover, researchers (e.g., Kunina-Habenicht et al., 2017; Sinharay & Haberman, 2009) asserted that the person applying the DCM should demonstrate the incremental validity of DCM estimates over and above simple subscores or overall scores estimated by simpler models. As such, given the principle of parsimony, use of DCMs is justifiable only if DCMs provide more reliable and valid information about examinees than do simpler models.

Relative Advantages and Limitations of DCMs

DCMs are considered statistically powerful diagnostic tools with the following notable advantages:

- DCMs can provide more reliable information about the examinee's relative standing on each attribute than can other diagnostic methods. This is primarily because DCMs assume discrete latent traits; thus, they measure the reliability at the categorical level and yield estimates of a smaller range compared to continuous constructs (Templin & Bradshaw, 2013).
- DCMs can provide a more reliable, statistically driven classification of examinees regardless of whether measures are norm-referenced or not. This is because DCMs classify examinees into discrete categories based on the PPM rather than cut scores (Min & He, 2021; Rupp et al., 2010). In contrast, in CTT or IRT context, the classification of examinees requires standard-setting procedures to derive a series of cut scores that are typically pre-specified relative to the population mean (Rupp et al., 2010). Such

procedures involve subjective decision-making and hence introduce additional sources of errors.

- DCMs allow for the modeling of complex data structures and potential attribute hierarchies by utilizing the Q-matrix. In effect, Q-matrix allows for the effective estimation of multiple skills using a relatively small set of items and sample sizes by limiting the number of parameters to be estimated.
- DCMs can inform efficient learning paths customized to individual needs based on the association between a given attribute pattern and expected gains in test scores (Rupp et al., 2010; Templin & Bradshaw, 2014). With increased emphasis on individualized support for underachieving students (National Joint Committee on Learning Disabilities [NJCLD], 2011) and renewed recognition that no two children take the same path to learning (Clay, 1998, 2002; Y. Goodman et al., 2005), such information is considered valuable, especially for early prevention and intervention.
- DCMs can provide detailed information about the model-data fit at attribute, attribute pattern, and test levels. This information helps further evaluate the validity of the classification (Rupp et al., 2010).

On the other hand, DCMs have several disadvantages as follows:

- DCMs require intensive computation and the use of iterative algorithms to obtain estimates of model parameters (Sen & Cohen, 2021).
- DCMs also lack user-friendly programs to estimate parameters although recent advances in R package development are likely to facilitate the use of DCMs in practice.

- DCMs have relatively large sample size requirements, albeit smaller than those of MIRT models, which makes their applications in small-scale classroom contexts almost impossible (Ravand & Baghaei, 2019).
- DCMs require the formulation of a correct Q-matrix. This critical element of DCMs poses a primary challenge to the application of DCM because it requires expertise in the cognitive domain of interest.
- DCM applications often report highly correlated attributes. Sessoms and Henson (2018) found that of DCM applications that computed attribute correlations, 90% reported multiple correlations exceeding .90. High attribute correlations indicate non-distinctive attribute estimates. As is the case with other diagnostic methods, the lack of attribute distinctiveness is largely attributable to the retrofitting approach in which DCMs are fitted to essentially unidimensional tests that were designed for non-diagnostic purposes.

Summary of Literature Review and Research Questions

Among various subscore estimation techniques, AIRT and MIRT approaches stood out as promising diagnostic tools providing more accurate and precise estimates when compared to raw subscores and theta estimates from UIRT. AIRT and MIRT approaches use ancillary information to improve the quality of subscores. Researchers have indicated that the degree of improvement largely depends on the reliability of the in-scale and out-of-scale items as well as inter-subdomain correlations. Concerning the relative advantages and disadvantages of these approaches, the MIRT approach is considered more flexible and theoretically appealing than the augmentation approach in that it explicitly models the multidimensionality of the test data incorporating the correlational structure of subscores into the estimation procedure to simultaneously estimate all the subscores. However, the flexibility of the MIRT approach comes with a price: its complexity

involves a heavier computational burden and more data demand for accurate and stable estimation of multiple latent abilities.

Although these statistically enhanced subscores improve the subscore reliability, they may not always provide the added value over and above the total score. Researchers who employed three added-value analysis techniques – PRMSE (Haberman, 2008), between- and within-person subscore reliability with profile analysis (Bulut et al., 2016), and criterion-related validity evidence of subscores (Davison et al., 2015) – have demonstrated that there is a delicate trade-off between improving subscore reliability and decreasing subscore orthogonality when subscores are statistically enhanced by borrowing strength from the rest of the test. Given that the subscore added value is the function of both reliability and distinctiveness of subscores, the utility of subscore estimation technique will depend on the extent to which the increased reliability it offers can offset the inevitable increase in inter-subdomain correlations (Feinberg & Wainer, 2014).

Researchers have indicated that DCMs have great potential as a diagnostic tool in that they can achieve higher levels of reliability for classifications, allow for modeling of complex data structure and potential attribute hierarchies, inform efficient learning paths to improve instruction and student learning and provide a detailed model-data fit information at different levels. Ironically, DCMs can afford more granularity in attributes and achieve greater accuracy and precision in estimation because they assume more coarse latent traits (i.e., discrete latent traits) than CTT or IRT (Rupp et al., 2010). However, in contrast to the abundance of research on the quality of subscores estimated from CTT and IRT, there has been a dearth of research evaluating the psychometric quality of DCM estimates, especially in comparison with other subscore estimates from IRT and CTT frameworks.

Therefore, in the present study, I aimed to compare DCMs with AIRT and MIRT under various statistical conditions with respect to classification accuracy, distinctiveness, and incremental validity evidence of subscores. As noted earlier, the reliability of DCM estimates is evaluated in terms of classification accuracy or consistency. Besides, it is not unusual that classification decisions (e.g., pass/fail; placement in some achievement level; or eligibility for remedial instruction) are made in CTT or IRT context as well by comparing examinees' continuous test scores to pre-specified cut-scores (e.g., Edwards & Vevea, 2006; Yao & Boughton, 2007). Therefore, for parallel comparisons, the reliability of subscores should be evaluated in terms of classification accuracy.

In addition, it is of interest to evaluate the extent to which point estimates (i.e., discrete mastery categories or continuous proficiency scores) at the subdomain level are correlated with each other and with those at the test level (i.e., overall scores). For example, if subdomain-level classifications are too highly correlated with overall classifications, classification decisions at a specific cut-score should be based on overall scores rather than subscores (Chin, 2011). Therefore, subscore distinctiveness should be evaluated using intercorrelations of subscores and correlations between subscores and overall scores based on both discrete mastery categories and continuous proficiency scores.

Lastly, as suggested by DCM researchers, it is imperative to examine the criterion-related validity evidence of DCM estimates to ensure that the diagnostic feedback from DCM supports the intended use of test scores. Moreover, if DCM estimates do not better predict the external criterion over and above the estimates from simpler approaches, the use of DCMs may not be justifiable given the principle of parsimony. Therefore, the criterion-related validity evidence of subscores should be evaluated relative to that of overall ability estimates from a simpler approach.

To achieve study goals, I conducted a simulation study manipulating a wide array of conditions found to affect reliability and added value of subscores in the previous studies. Specific research questions are as follows:

1. How do AIRT, MIRT, and DCMs compare in terms of the classification accuracy (based on discrete mastery categories), at the varying levels of subscale length, item difficulty distribution, and inter-subscore correlations as well as across conditions?
2. How do AIRT, MIRT, and DCMs compare in terms of the extent to which subscores (based on both discrete mastery categories and continuous proficiency scores) are distinct from each other and from overall scores, at the varying levels of subscale length, item difficulty distribution, and inter-subscore correlations as well as across conditions?
3. How do AIRT, MIRT, and DCMs compare in terms of the extent to which subscores (based on continuous proficiency scores) contribute to the prediction of the external criterion over and above overall scores estimated from the unidimensional IRT (UIRT) approach, at the varying levels of subscale length, item difficulty distribution, and inter-subscore correlations as well as across conditions?

Chapter 3

Methodology

The Monte Carlo simulation was conducted to answer research questions presented in the previous section. To illustrate the study design in detail, this section is organized into five topics: (a) simulation condition, (b) data generation, (c) subscore estimation, (d) evaluation criteria, and (e) data analysis.

Simulation Condition

This section is organized into two parts, fixed study design elements and manipulated study design elements. Fixed study design elements refer to simulation conditions that are invariant across all replications, whereas manipulated study design elements refer to primary independent variables employed in this study.

Fixed Study Design Elements

This section describes the following three fixed study design elements: (a) data structure, (b) sample sizes, and (c) the number of subdomains.

Data Structure. The assessment in this study was generated based on a simple structure (i.e., each item measured only one attribute). Although a simple structure was rather a restricted condition, it has been preferred over the complex structure in most multidimensional assessments due to its superb estimation and classification accuracy and interpretation simplicity (Feinberg & Wainer, 2014; Madison & Bradshaw, 2015).

Sample Sizes. Researchers (Sinharay et al., 2010; X. Wang et al., 2019) found that sample sizes had no impact on the proportion of subscores that have added values. Other researchers also reported a negligible impact of sample sizes on ability estimates in unidimensional and multidimensional IRT (e.g., de la Torre & Song, 2009; Yao & Boughton, 2007) and on classification accuracy in DCM (e.g., de la Torre et al., 2010; Kunina-Habenicht et

al., 2012; Sen & Cohen, 2021). However, sample sizes affect statistical power for detecting significant effects; therefore, adequate sample sizes are needed when significance tests are involved as is the case for validity analysis. In general, larger sample sizes are required for stable and accurate estimation of model parameters and abilities as the number of subdomains measured by each item and the whole test increases and more complex models are employed. Considering the study design employed in this study, the sample size was fixed at 2,000 examinees per data file. Based on previous simulation studies (e.g., Edwards & Vevea, 2006; Kunina-Habenicht et al., 2012; Yao & Boughton, 2007), the selected sample size was considered large enough to allow for adequate statistical power for detecting significant effects and to avoid potential convergence issues involved in the estimation of IRT and DCM parameters.

The Number of Subdomains. For the sake of simplicity, the number of subdomains was constrained to two in this study. Researchers (e.g., Bulut et al., 2016; Edwards & Vevea, 2006; Sinharay, 2010; X. Wang et al., 2019) reported that the number of subdomains had no or negligible impact on reliability and/or added value of subscores estimated from CTT or IRT framework. However, it should be noted that like sample sizes, the number of subdomains influences the results of significance tests through degrees of freedom. The insignificant impact of the number of subdomains reported in those studies stems from the fact that most of them did not involve significance tests.

Manipulated Study Design Elements

Simulation conditions were manipulated for the following five factors: (a) subscore estimation methods (DM; four levels), (b) subscale length (L; four levels), (c) distributions of item difficulties (D; three levels), (d) inter-subdomain correlations (R; four levels), and (e) magnitude of criterion validity coefficients (C; two levels). All these manipulated factors (MF) except the subscore estimation methods and criterion validity coefficients were fully crossed

during data generation yielding a total of 4,800 unique data sets. For each data file, person abilities (i.e., subscores or overall scores) were estimated by four methods (DM) and criterion variables were generated based on two levels of criterion validity coefficients (C). All the manipulated factors are summarized in Table 3.1.

Subscore Estimation Methods. To compare performances of diagnostic methods (DM) with respect to classification accuracy and incremental criterion-related validity evidence, I employed three subscore estimation methods (i.e., AIRT, MIRT, and DCMs) plus a unidimensional IRT (UIRT) approach. The UIRT was utilized to obtain an overall ability estimate to which subscore estimates from three diagnostic methods were compared to evaluate their distinctiveness and added values. It should be noted that these methods were manipulated during data analyses but not during data generation. Thus, each of generated data files was analyzed four times, each time with a different estimation method. A detailed description of these methods is provided in a later section for score estimation.

Subscale Length. The subscale length (L) is known to have a substantial effect on the reliability and added value of subscores. Researchers (e.g., Puhan et al., 2010; Sinharay, 2010) suggested that at least 20 multiple-choice items per subtest are required to obtain moderately high reliability in a realistic situation where inter-subdomain correlations are .70 or higher. In addition, the majority of operational tests were found to have 10 to 30 items per subscale although average subdomain test lengths ranged from 11 to 69 items (Sinharay, 2010). By contrast, researchers of DCMs (e.g., Hartz, 2002; Jang, 2009) suggested a minimum of three to five items per attribute for the appropriate level of estimation and classification accuracy. Taking these findings into consideration, the effect of the subscale length was explored using the following four levels of the subscale length: 5, 10, 20, and 30 items per subscale. As an equal number of items was assigned to each of the two subscales, the total number of test items ranged from 10 to 60.

Distribution of Item Difficulties. Researchers suggested that item characteristics such as level and variability of item difficulty affected reliability and added value of subscores. As such, Gulliksen (1945) and Symonds (1928) theoretically proved that extreme values (too hard/easy) and larger variability of item difficulty reduced score variance thus leading to lower test score reliability. In light of their notions, Wang et al. (2019) conducted a simulation study to investigate how properties of the item difficulty distribution influence reliability and added value of subscores. The authors found that the subscore reliability and the proportion of subscores of added value tended to decrease with larger variability of item difficulty and mismatch between average item difficulty and average examinee ability when the subscale length was relatively short. Researchers also found a similar detrimental effect of extreme item difficulty on the classification accuracy of DCMs. In an applied analysis, Jang (2009) found that too easy or too hard items had poor diagnostic capacity and thus led to lower classification accuracy with other things being equal.

Therefore, to explore effects of item difficulty distribution (D) on classification accuracy, three levels of item difficulty distributions were created as follows: (a) baseline (BL): $b \sim N(0, 1)$; (b) high variance (HV): $b \sim N(0, 1.5^2)$; and (c) high mean (HM): $b \sim N(1, 1)$. The high variance condition was characterized by the standard deviation (*SD*) 1.5 times larger than that of the baseline condition, whereas the high mean condition featured hard tests with the mean difficulty increased by one *SD* compared to the baseline condition.

Inter-Subdomain Correlations. Researchers (e.g., Davison et al., 2015; de la Torre et al., 2011; de la Torre & Patz, 2005; Edwards & Vevea, 2006; Fu & Qu, 2018; Sinharay, 2010; Stone et al., 2010; Thissen & Wainer, 2001) have demonstrated that inter-subdomain correlations are closely related to the value of reporting subscores. In an extensive simulation study to examine factors affecting the added value of subscores, Sinharay (2010) suggested an upper limit

of average disattenuated inter-subdomain correlations for subscores to be of added value: .70 for subscales with only 10 items and .80 for subscales with 20 items. He also found that with correlations of .90 or higher, subscores rarely had added value regardless of test length although disattenuated inter-subdomain correlations typically varied between .7 and .9. Similarly, Lyrén (2009) and McPeek et al. (1976) noted that average inter-subdomain correlations should be less than .90 for subscores to add value over the total score.

To examine the effect of the magnitude of inter-subdomain correlations, a wide range of true inter-subscore correlations (R) were included in this study: .30, .50, .70, and .80. The high correlations of .70 and .80 were included to reflect the cut values reported by Sinharay (2010). Under these high correlation conditions, subscores were anticipated to have added value under certain conditions only. In contrast, low and moderate correlations of .30 and .50, albeit less common, were included to explore how other simulation conditions interact under a more ideal context of multidimensionality. Note that the true inter-subscore correlations are similar to disattenuated correlations (i.e., correlations corrected for unreliability) (Sinharay, 2010).

Magnitude of Criterion Validity Coefficients. The criterion variable was created to explore the extent to which subscore estimates had incremental validity evidence over and above the overall ability estimate with respect to the prediction of the external criterion variable. The criterion variable was generated based on its correlation with true person parameters, θ_1 and θ_2. Note that the square of this correlation corresponds to the proportion of variation in the criterion variable that is explained by true ability estimates (shared variance). Like diagnostic methods (DM), the magnitude of criterion validity coefficients was not manipulated during the data generation and thus did not affect simulated item scores.

To explore the impact of the magnitude of criterion validity coefficients (C) on incremental validity evidence of subscores, the criterion variable was generated based on two

levels of population correlations: $\rho = .80$ and $\rho = .50$. The population correlation of .50 was selected to reflect moderate correlations (Cohen, 1988) as well as average correlations between subscores and criterion variables reported in the empirical study by Kunina-Habenicht et al. (2017). In contrast, the population correlation of .80 was selected to reflect high correlations between subscores and criterion variables. Each of these correlations corresponded to the designated criterion validity evidence of 64% shared variance (i.e., high criterion validity, HCV) and 25% shared variance (i.e., low criterion validity; LCV), respectively. The steps to generate criterion variables were as follows: (a) two random variables having the designated population correlation (i.e., $\rho = .80$ or $\rho = .50$) with each of two true examinee abilities (i.e., θ_1 and θ_2) were generated, and (b) these two random variables were summed up to yield a criterion variable. The same procedure was repeated twice to generate criterion variables based on HCV and LCV conditions in each cell. Note that in this study criterion variables were generated to have a similar relationship with two subscores. Score values for criterion variables with high validity coefficients ranged from -8.2 to 7.9 with $M = 0$ and $SD = 1.7$, whereas score values for criterion variables with low validity coefficients ranged from -6.9 to 7.1 with $M = 0$ and $SD = 1.5$. In addition, the average correlation between criterion variables and true examinee abilities was .76 ($SD = .06$) under high validity condition and at .52 ($SD = .06$) under low validity condition.

Table 3.1

Manipulated Study Design Elements

MF	Level
Diagnostic methods (DM)	AIRT
	MIRT
	ACDM
	UIRT
Subscale length (L)	5 items per subscale
	10 items per subscale
	20 items per subscale
	30 items per subscale
Distribution of item difficulty (D)	BL: b ~ N (0, 1)
	HV: b ~ N (0, 1.5^2)
	HM: b ~ N (1, 1)
Inter-subscale correlations (R)	.30
	.50
	.70
	.80
Magnitude of criterion validity coefficients (C)	HCV: 64 % criterion validity
	LCV: 25 % criterion validity

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, UIRT = unidimensional IRT, BL = baseline, HV = high variability, HM = high mean, HCV = high criterion validity, LCV = low criterion validity.

Data Generation

In what follows, data-generating parameters (person and item parameters) and model and overall data generation procedure are illustrated.

Person Parameters

Person parameters for two subdomains were drawn from multivariate normal distributions with $M = 0$, $SD = 1$, and appropriate covariance to account for the intercorrelations

of true examinee abilities corresponding to the conditions described in the previous section. That is, $\boldsymbol{\theta} \sim N\ (\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{0} = (0, 0)'$ and $\boldsymbol{\Sigma}$ is a variance-covariance matrix for two latent abilities. Given that $SD = 1$, diagonals of $\boldsymbol{\Sigma}$ are 1 and off-diagonals corresponded to the inter-subdomain correlations (i.e., .30, .50, .70, or .80) manipulated in each simulation condition. Values of true person parameters were consistent across simulation conditions ranging from -5.0 to 4.7 with $M = 0$ and $SD = 1.0$.

Item Parameters

Item discrimination parameters were randomly drawn from 62 discrimination parameters estimated for the National Assessment of Educational Progress (NAEP) Grade 4 reading assessment. Values of true discrimination parameters were consistent across levels of manipulated factors ranging from 0.27 to 2.30 with $M = 1.02$ and $SD = 0.39$. On the other hand, given that the distribution of item difficulty parameters (b) was a manipulated factor in this study, the difficulty parameters were drawn from a univariate normal distribution with M and SD corresponding to the conditions described in the previous section. In this study, the item difficulty distribution at baseline was aligned with the examinee ability distribution, whereas item difficulty distributions at high variability and high mean conditions were deliberately mismatched with the examinee ability distribution. True difficulty parameters ranged from -5.8 to 6.0 with $M = 0.3$ and $SD = 1.2$ across simulation conditions but as expected they varied depending on the manipulated conditions of item difficulty distribution. They ranged from -3.9 to 4.0 with $M = 0.0$ and $SD = 1.0$ at baseline, they ranged from -5.8 to 6.1 with $M = 0.0$ and $SD = 1.5$ at high variability condition, and they ranged from -2.9 to 5.0 with $M = 1.0$ and $SD = 1.0$ at high mean condition.

Data-Generating Model

Binary item scores were simulated based on a two-dimensional two-parameter normal ogive model (2PNO). Its IRF is identical to Equation 19 with a scaling constant of 1.7. Due to the

challenge in the estimation of non-compensatory MIRT models, the compensatory model was selected over the non-compensatory model. In addition, the choice of 2PNO over the three-parameter normal ogive (3PNO) model was based on prior findings that the estimation of model parameters was negatively affected by the presence of the lower-asymptote parameter in some MIRT estimation methods when items had a simple structure (Finch, 2010; Tate, 2003).

Data Generation Procedure

For each replication in the given simulation condition, the person parameter $\boldsymbol{\theta}$ for each of the 2,000 examinees was drawn from the multivariate normal distribution $N\ (\mathbf{0}, \boldsymbol{\Sigma})$ with the corresponding inter-subdomain correlations. Item parameters for two subdomains (i.e., $L \times 2$) were also drawn as described above. Then, binary scores of each item were simulated for each examinee based on two-dimensional 2PNO by plugging withdrawn person and item parameters in Equation 19. For each simulated data set, evaluation criteria of interest were calculated and saved for further analyses. Given that Harwell et al. (1996) recommended a minimum of 25 replications for Monte Carlo simulation studies based on IRT models, a hundred replications were carried out for each simulation condition. Both person and item parameters were redrawn in each replication using a unique starting seed, thus resulting in 4,800 unique response data files for 48 simulation conditions. Note that the same 100 starting seeds were used in each cell, so the mean difference between cells was only attributable to different simulation conditions rather than different starting seeds.

Subscore Estimation

Using simulated item scores, subscores were estimated from three diagnostic methods (i.e., AIRT, MIRT, and DCMs), whereas overall scores were estimated by UIRT. This section is organized into three parts: (a) estimators, (b) estimation of mastery categories, and (c) estimation methods employed in this study.

Estimators

Model parameters were estimated with the standard EM algorithm which is an iterative method for finding maximum likelihood estimates of parameters for probability models. The EM algorithm is effective for the MIRT models with up to three dimensions (Chalmers, 2012). Each iteration of the EM algorithm consists of two steps: the E step (expectation) and the M step (maximization). The EM algorithm is applied to estimate the parameters for each item individually, and then the iteration process is repeated until certain convergence criteria (e.g., the changes of likelihood function values and all item parameter estimates are smaller than preselected values) are met.

On the other hand, subscores from IRT-based methods were estimated with MAP. As Bayesian estimator, MAP solves the likelihood function by maximizing a posterior distribution based on prior information and computes the mode of the posterior distribution as the final estimate of the latent trait. According to Bulut (2013), MAP is preferable over other estimators such as ML, EAP, and MCMC, considering the following aspects: (a) as Bayesian estimator, MAP allows for incorporation of collateral information, thus leading to more precise subscore estimates (P. H. Chen, 2006; de la Torre et al., 2011; Yao, 2010); (b) EAP, another Bayesian estimator, computes the mean of the posterior distribution to estimate the latent trait, which may however lead to biased estimates of the scores (Wainer & Thissen, 1987); (c) for high-dimensional tests, MAP requires far less computation time than EAP; (d) MAP allows for estimating scores from all possible response patterns (e.g., all-correct response pattern, all-incorrect response pattern), which is not possible in ML because ML estimates the subscores by maximizing the likelihood of an examinee's item responses (Embretson & Reise, 2000); (e) MCMC is computationally intensive and inconvenient for testing programs with large sample sizes (de la Torre & Patz, 2005; Yao & Boughton, 2007).

In contrast, DCM attributes were estimated using EAP. According to Rupp et al. (2010), EAP is preferable over MAP when the individual attribute is of interest given that EAP provides direct probability estimates for each attribute separately. Authors noted that MAP estimates could be sometimes difficult to interpret without such information.

Estimation of Mastery Categories

In DCMs, the mastery status of each attribute was computed based on the PPM – also known as marginal probabilities of mastery for individual attributes – for the EAP estimate of attributes. As stated earlier, examinees with the PPM equal to .50 or greater were considered to have mastered the given attribute. On the other hand, the continuous subscores estimated from IRT-based methods were further dichotomized based on cut-scores to create discrete mastery categories. In the present study, the cut-score was set at $\theta = 0$ across conditions for the following reasons.

Firstly, to compare classification accuracy between diagnostic methods, the cut-score for the continuous models should be set at a level that is comparable to the mastery level defined by DCMs (Chin, 2011). Preliminary results indicated that all diagnostic methods produced a comparable base rate of mastery when examinees were classified at $\theta = 0$ when average item difficulty was also 0. Secondly, it is often desired to set cut-scores at average item difficulty where test information is maximized. In the present study, two of the three item difficulty conditions had a mean item difficulty of 0 which also coincided with average examinee ability. Lastly, DCMs are designed to provide criterion-referenced rather than norm-referenced interpretation. Thus, for parallel comparisons of diagnostic methods, classifications should be made based on the same cut-scores across conditions given that this study mimicked the situation where the same group of examinees takes multiple tests measuring the same construct but having different difficulty distributions.

Estimation Methods

UIRT. The unidimensional 2PNO model (Equation 18 with the scaling constant of 1.7) was used to estimate a single dimension that represented an overall ability. As noted earlier, multidimensional estimates from three diagnostic models were compared with the UIRT estimate to examine their distinctiveness and added value over and above the overall ability estimate. Hereafter, the overall score refers to estimates from UIRT. The estimation of IRT models was carried out using the R package *mirt* (Chalmers, 2012).

AIRT. To obtain augmented theta estimates, each dimension was first estimated separately based on the unidimensional 2PNO model. Then, initial MAP estimates were corrected using Equation 15 for their tendency to shrink toward the population mean before being used as predictors in the regression model (Equation 12).

MIRT. The two-dimensional 2PNO model (Equation 19 with the scaling constant of 1.7) was used to simultaneously estimate subscores for two subdomains incorporating their correlational structure into the estimation process.

DCM. In accordance with compensatory MIRT models employed for data generation and estimation, ACDM (Equation 27) was adopted to calibrate attributes from the DCM approach. As a compensatory model, ACDM is equivalent to a common parameterization of the saturated CDM (e.g., G-DINA, LCDM) with main effects only (i.e., no interaction effects). The choice of ACDM over non-compensatory DCMs such as DINA was also based on the prior findings that interaction effect parameters were often estimated unreliably except for very large sample sizes and were also of relatively minor importance for predicting observed data patterns in the presence of main effect parameters in the model (Kunina-Habenicht et al., 2012, 2017). The estimation of ACDM was carried out by the R package *GDINA* (Ma & de la Torre, 2020).

Evaluation Criteria

Convergence

To evaluate the extent to which the given diagnostic method was practically feasible for each condition, the number of iterations in each cell that failed to converge was computed for each estimation method. In this study, tolerance of .0001 was used as convergence criteria during the estimation. The replications that didn't properly converge were excluded from further analyses.

Model-Data Fit

Acceptable fit is a prerequisite to examining model results (R. Liu et al., 2018). If the fit of the given model is not satisfactory, it is likely to result in inaccurate and imprecise subscore estimates. Thus, as the first step, the model-data fit was reviewed to evaluate whether each diagnostic model adequately fitted data under each condition. As a primary interest of this study was model comparisons rather than individual items or persons, the model-data fit was inspected only at the test level using absolute and relative fit indices. Absolute fit indices designate the adequacy of fit of a single model, whereas relative fit indices designate whether there is a significant difference in the fit of competing models.

The choice of fit indices was based on their popularity in studies involving IRT and DCM models (e.g., Kunina-Habenicht et al., 2017; R. Liu et al., 2018; Min & He, 2021; Nye et al., 2020). The absolute fit of models was evaluated using the standardized root mean square residual (SRMSR; Maydeu-Olivares, 2013). The model with SRMSR smaller than or equal to .05 is considered well-fitting with a substantively negligible amount of misfit. Thus, the number of iterations in each cell having SRMSR greater than .05 was computed for each estimation method. On the other hand, the relative fit of models was evaluated using Akaike Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC; Schwarz, 1978) for which the

smaller values indicated a better fit. Note that BIC imposes a large penalty for highly parameterized models.

Classification Accuracy

To compare the classification performance of diagnostic methods (research question 1), the classification accuracy was evaluated using the following four measures of agreement: (a) correct classification rate (CCR), (b) kappa (κ), (c) sensitivity (Sens), and (d) specificity (Spec). These indices were computed using 2 × 2 contingency tables (Table 2.1) – also known as confusion matrix of classification – where proportions of the examinee population were cross-classified by their true mastery categories and their estimated mastery categories. Note that with reference to signal detection theory (Green & Swets, 1966), p_{00}, p_{01}, p_{10}, and p_{11} in Table 2.1 also represent the proportion of correct rejection, false alarm, miss, and hit, respectively.

In this study, two subscales were generated with equal length based on the same parameter distribution specifications, and mastery classifications of each subscale were based on the same cut-score. Given such symmetry of the two subscales, the difference in their classification accuracy was negligible. Thus, the classification accuracy was examined by combining confusion matrices of two subscales; for each replication, cell proportions in Table 2.1 were computed relative to total counts of 4,000 (2 subdomains × 2,000 examinees) rather than 2,000. According to Chin (2011), this collective approach is also advantageous in that it reduces unnecessary repetitions in data analyses and prevents Type I error rate inflation.

Mastery Rate. The mastery rate was defined as the proportion of examinees classified as maters. The mastery rate estimated for each diagnostic method was compared to the true mastery rate. As precedent for classification accuracy, the discrepancy between the estimated and the true mastery rate indicated the untrustworthiness of the given diagnostic method for mastery classifications.

Correct Classification Rate. The correct classification rate (CCR), also known as the proportion correctly classified (P_c; Clogg, 1995; de la Torre & Douglas, 2004), is defined as the proportion of examinees within a sample whose estimated classification is accurate. It is equivalent to the attribute-level classification accuracy shown in Equation 30 (i.e., CCR = $p_{00} + p_{11}$). Despite its straightforward interpretation, CCR is limited in that the number of matches merely due to chances increases when base rates of mastery approach the extremes (i.e., 0 or 1) (Chin, 2011).

Cohen's Kappa. Cohen's (1960) kappa (κ) is preferable over CCR in that it considers base rates of mastery categories and thus corrects for the artificial chance agreement in the CCR . The κ was computed by Equation 34.

Sensitivity. Neither CCR nor κ informs the specific types of accuracy. To further differentiate hit from correct rejection, the sensitivity (Sens) was calculated by Equation 35.

Specificity. As a counterpart of sensitivity, specificity (Spec) was computed by Equation 36.

Subscore Distinctiveness

Agreement measures described above indicate the extent to which subscores are accurate and reliable with respect to classifications. For subscores to provide meaningful diagnostic information, they must be also distinct from each other and from the overall scores (Sinharay, 2014). Therefore, to assess subscore distinctiveness (research question 2), inter-subscore correlations of each diagnostic method as well as correlations between subscores and overall score estimates were computed. As the present study utilized both binary mastery status and continuous proficiency scores, tetrachoric and zero-order Pearson correlations were obtained for each pair of scores. As noted earlier, PPM served as proficiency scores for DCMs.

The proportion of Variation in Criterion Explained by Subscores vs. Overall Score

To assess the extent to which diagnostic subscores add values over and above overall scores concerning the prediction of a criterion variable (research question 3), a series of hierarchical/block-wise regression models were fit. First, overall scores alone (Model 1, M_1) were added as the first block of predictors. Then, two subscores of each diagnostic method were added as a second block of predictors (Model 2, M_2). The hierarchical regression is often used for model comparison of nested regression models. Following the standard practice for regression models, the incremental validity of each diagnostic method was reported using the change in adjusted R^2 between Model 1 and Model 2; if the difference in adjusted R^2 between Model 1 and Model 2 is significant, it can be said that the added predictors in Model 2 (i.e., two subscores) explain the criterion variable over and above the predictors in Model 1 (i.e., overall scores) (Kunina-Habenicht et al., 2017). Note that the adjusted R^2 takes into account the number of predictors in the model. Also, to explore the impact of the magnitude of criterion validity coefficients on subscore contribution over and above overall scores, R^2 of two regression models were obtained separately for criterion variables with high validity and low validity coefficients.

Data Analysis

To check the feasibility of each estimation method, their convergence and model-data fit was first evaluated using indices described earlier. Then, in a separate run for each dependent variable (DV), a mixed analysis of variance (ANOVA) was performed to examine the effect of simulation conditions on the classification accuracy and the proportion explained in the criterion variability (research questions 1 and 3).

Mixed Analysis of Variance

Each mixed ANOVA comprised one within-subject factor (diagnostic methods, DM) and three between-subject factors (subscale length, L; difficulty distribution, D; and inter-subscore correlations, R). The dependent variables (DVs) included CCR, κ, Sens, and Spec as well as

adjusted R^2 of M_1 and M_2 under HCV and LCV conditions. Note that mixed ANOVA of classification accuracy measures was conducted on three diagnostic methods (i.e., within-subject factor comprised three levels, AIRT, MIRT, and ACDM). In contrast, mixed ANOVA of adjusted R^2 was conducted on all four estimation methods (i.e., UIRT, AIRT, MIRT, and ACDM) to evaluate the incremental validity evidence of diagnostic methods over and above the unidimensional approach (research question 3). As the within-subject factor had more than two levels, the sphericity assumption of mixed ANOVA was evaluated with the Mauchly test. Whenever the Mauchly test indicated the violation of the sphericity assumption, Greenhouse-Geisser correction (Greenhouse & Geisser, 1959) was applied to the within-subject factor.

Post-Hoc Analyses

Post-hoc analyses were performed for significant effects of mixed ANOVA. To prevent following up on too many effects that were statistically significant but practically not meaningful, significant effects were identified based on the effect size and the statistical significance. The effect size of mixed ANOVA was measured by generalized eta squared (η_G^2) (Olejnik & Algina, 2003). The generalized eta squared is preferable over eta squared or partial eta squared given that it allows for comparison of effect sizes between studies not only with similar experimental designs but also with different experimental designs (Olejnik & Algina, 2003). According to Cohen (1988, p. 25), for an effect "likely to be visible to the naked eye of a careful observer," the size of the effect should be at least medium. Cohen (1988) has also provided guidelines to define small (eta squared = .01), medium (eta squared = .06), and large (eta squared = .14), effects. According to Thompson (2007), generalized eta squared can be evaluated based on the guidelines provided by Cohen although it is preferable to relate the effect size to other effects in the literature. Thus, in the following sections, "significant" effects refer to those having η_G^2 equal to or greater than .06 along with significant *p*-values.

To follow up on significant main effects, pairwise comparisons of the estimated marginal means were conducted. Note that effect sizes were not available for pairwise comparisons; therefore, their significance was evaluated solely by *p*-values. In addition, significant two-way interactions were further examined by (a) simple main effects (i.e., the main effect of the first variable at fixed levels of the second variable) and (b) simple comparisons (i.e., pairwise comparisons between means of the first variable at fixed levels of the second variable). The significance levels of post hoc tests were adjusted by the Bonferroni correction such that the nominal significance level was divided by the number of comparisons to keep the family-wise Type I error rate smaller than .05. Although the Bonferroni correction is more conservative relative to other available correction procedures, it is widely used owing to its straightforward application. As with data generation, statistical analyses were performed by R version 4.2.1 (R Core Team, 2021).

Chapter 4

Results

This chapter comprises five sections containing the results of statistical analyses. The first two sections include results related to convergence and model-data fit of each method. The remaining three sections are organized in the order of results relevant to each of the three research questions.

Convergence

Table 4.1 contains the number of iterations in which each estimation method failed to converge. In general, the convergence rate of each method was very high. ACDM successfully converged in all 4,800 iterations, and MIRT failed to converge in only one iteration. In contrast, UIRT and AIRT failed to converge in 7 and 12 iterations, respectively. The relatively poorer convergence of AIRT is attributable to the fact that each subscore of AIRT was estimated separately based on the unidimensional method before being augmented by the regression approach; the convergence of AIRT was flagged as successful only if two separate calibrations of subscores met prespecified tolerance criteria. As a result of unsuccessful convergences, a total of 15 iterations out of 4,800 were removed from further analyses.

Table 4.1

Counts of Unsuccessful Iterations for Each Method

L	D	R	UIRT	AIRT	MIRT	ACDM
5	BL	.30	0	0	1	0
10	HV	.50	0	2	0	0
10	HV	.80	0	1	0	0
20	HV	.30	0	2	0	0
20	HV	.50	0	1	0	0
20	HV	.70	2	2	0	0
20	HV	.80	1	2	0	0
20	HM	.80	1	0	0	0
30	HV	.30	1	0	0	0
30	HV	.70	1	1	0	0
30	HV	.80	1	1	0	0
	All		7	12	1	0

Note. UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Model-Data Fit

SRMSR

AIRT and MIRT had SRMSR equal to or smaller than .05 (i.e., good fit) in all iterations, whereas UIRT and ACDM had SRMSR greater than .05 in 794 and 3 iterations, respectively (Table 4.2). Both UIRT and ACDM yielded better absolute fit when subscales were shorter and item difficulty changed from BL to HM and HV. However, as inter-subscale correlations increased, the absolute fit of UIRT improved, whereas that of ACDM deteriorated.

Table 4.2

Counts of Iterations Where Each Method had SRMSR Greater than .05

L	D	R	UIRT	AIRT	MIRT	ACDM
5	BL	.30	53	0	0	0
5	BL	.50	4	0	0	0
5	HV	.30	21	0	0	0
5	HM	.30	37	0	0	0
5	HM	.50	3	0	0	0
10	BL	.30	82	0	0	0
10	BL	.50	2	0	0	0
10	HV	.30	44	0	0	0
10	HM	.30	63	0	0	0
10	HM	.50	1	0	0	0
20	BL	.30	97	0	0	0
20	BL	.50	1	0	0	0
20	HV	.30	52	0	0	0
20	HM	.30	80	0	0	0
30	BL	.30	99	0	0	0
30	BL	.80	0	0	0	2
30	HV	.30	68	0	0	0
30	HM	.30	87	0	0	0
30	HM	.80	0	0	0	1
	All		794	0	0	3

Note. L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Information Criteria

Table 4.3 contains average information criteria by the level of each manipulated factor. Similar to the absolute model fit results, on average MIRT (M = 70,590 for AIC and M = 70,960 for BIC) and AIRT (M = 70,994 for AIC and M = 71,357 for BIC) had the best relative fit among

four estimation methods. Note that because two subscores of AIRT were estimated separately, AIC and BIC of AIRT were obtained by summing information criteria from two calibrations using a unidimensional approach. On average, UIRT (M = 71,492 for AIC, M = 71,856 for BIC) and ACDM (M = 71,498 for AIC, M = 71,879 for BIC) resulted in comparable information criteria that was higher than that of AIRT (M = 70,994 for AIC, M = 71,357 for BIC) and MIRT (M = 70,590 for AIC, M = 70,960 for BIC). Like SRMSR, both AIC and BIC improved as subscale length decreased, and item difficulty distribution changed from BL to HM and HV. In addition, all methods except AIRT yielded better relative fit as inter-subscore correlations increased.

In general, these results of convergence and model-data fit support the adequacy of three diagnostic models although DCMs yielded poorer fit indices relative to IRT-based methods. The poorer model-data fit of ACDM is primarily attributable to the fact that DCMs are based on assumptions that are different from those of the data generating model, MIRT. Whereas MIRT assumes continuous latent traits and nonzero variability of item difficulty, ACDM assumes discrete latent traits and equal difficulty across items (i.e., attribute homogeneity) with the primary purpose of classifying examinees according to multiple attributes. Researchers of DCMs claimed that these features enable DCMs to yield the reliability as high as IRT models with far fewer items per dimension (Templin & Bradshaw, 2013). Congruent with such strength, ACDM sometimes yielded the lowest AIC among four estimation methods when subscales were short (i.e., 5 items per subscale) and inter-subscale correlations were low (i.e., .30 and .50) although MIRT yielded the lowest relative fit indices in most of the cells.

Table 4.3

Average Information Criteria by Manipulated Factors

MF	Level	N	AIC			
			UIRT	AIRT	MIRT	ACDM
L	5	1199	22455	22531	22373	22408
	10	1197	44580	44551	44235	44476
	20	1192	88071	87465	86951	88039
	30	1197	131013	129578	128950	131222
D	BL	1599	75926	75332	74903	75918
	HV	1587	68912	68478	68085	68957
	HM	1599	69619	69152	68764	69601
R	.30	1196	72639	70999	70915	71708
	.50	1197	71903	71057	70807	71660
	.70	1197	70934	70953	70418	71372
	.80	1195	70492	70965	70220	71253
All		4785	71492	70994	70590	71498
MF	Level	N	BIC			
			UIRT	AIRT	MIRT	ACDM
L	5	1199	22567	22643	22491	22537
	10	1197	44804	44775	44465	44717
	20	1192	88519	87913	87405	88504
	30	1197	131685	130250	129628	131910
D	BL	1599	76291	75696	75273	76299
	HV	1587	69275	68842	68454	69337
	HM	1599	69983	69516	69133	69982
R	.30	1196	73003	71363	71284	72088
	.50	1197	72267	71421	71177	72041
	.70	1197	71298	71317	70787	71752
	.80	1195	70856	71328	70590	71634
All		4785	71856	71357	70960	71879

Note. MF = manipulated factors, AIC = Akaike information criterion, BIC = Bayesian information criterion, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Results for Research Question 1

This section contains the results of mixed ANOVA analyses performed to explore the comparative classification accuracy of three diagnostic methods under various conditions. This section is organized by the type of DVs. The first part presents mastery rates yielded from each diagnostic method along with true mastery rates given that further comparisons of classification accuracy build on these base rates of mastery. The subsequent four parts present results of mixed ANOVA conducted on each of the four agreement measures. Note that *p*-values reported for post-hoc analyses reflect Bonferroni adjustment (i.e., adjusted *p*-values).

Mastery Rate

True mastery rates were always .50 given that examinee ability and cut-scores were held constant across conditions (Table 4.4). All diagnostic methods tended to accurately estimate true mastery rates when average item difficulty was well aligned with average examinee ability and cut-scores. However, when the test was harder relative to examinee ability (i.e., at HM), ACDM substantially underestimated true mastery rates, whereas IRT-based methods estimated true mastery rates only with slight deviations (Figure 4.1).

Table 4.4

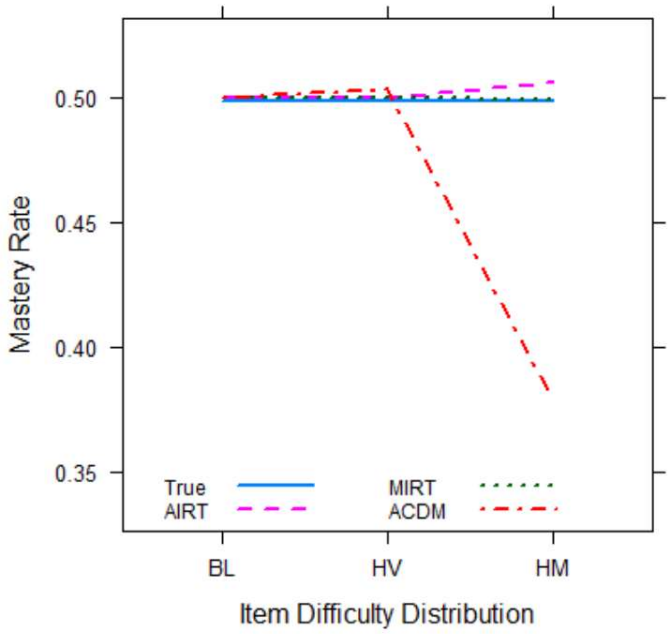
Average True and Estimated Mastery Rates by Manipulated Factors

MF	Level	N	True	AIRT	MIRT	ACDM
L	5	1199	.50	.51	.50	.45
	10	1197	.50	.50	.50	.46
	20	1192	.50	.50	.50	.46
	30	1197	.50	.50	.50	.47
D	BL	1599	.50	.50	.50	.50
	HV	1587	.50	.50	.50	.50
	HM	1599	.50	.51	.50	.38
R	.30	1196	.50	.50	.50	.46
	.50	1197	.50	.50	.50	.46
	.70	1197	.50	.50	.50	.46
	.80	1195	.50	.50	.50	.46
All		4785	.50	.50	.50	.46

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Figure 4.1

Average Mastery Rates by Item Difficulty Distribution



Note. AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Correct Classification Rates

Descriptive. As implied by base rates of mastery, CCR of AIRT (M = .827) and MIRT (M = .827) was comparable and slightly higher than that of ACDM (M = .815) across conditions. In addition, CCR increased with subscale length and inter-subscale correlations (Table 4.5). Concerning item difficulty distribution, CCR was highest at BL followed by HV and HM. Consistent with results of mastery rates, the difference between IRT-based methods versus ACDM was relatively more pronounced at HM although all three methods tended to yield comparable CCR at each level of item difficulty distribution.

Table 4.5

Correct Classification Rates by Manipulated Factors

MF	Level	N	AIRT	MIRT	ACDM
L	5	1199	.76	.76	.75
	10	1197	.81	.81	.80
	20	1192	.86	.86	.85
	30	1197	.88	.88	.87
D	BL	1599	.84	.84	.83
	HV	1587	.82	.82	.82
	HM	1599	.82	.82	.80
R	.30	1196	.82	.82	.81
	.50	1197	.82	.82	.81
	.70	1197	.83	.83	.82
	.80	1195	.84	.84	.83
All		4785	.83	.83	.82

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Main Effects. Table 4.6 contains mixed ANOVA results of CCR with Greenhouse-Geisser correction. Based on *p*-value ($p < .05$) and effect size criteria ($\eta_G^2 > .06$), the main effect of diagnostic methods (DM) was significant $F_{(1.07, 5085.44)} = 6744.81$, $p < .001$, $\eta_G^2 = .12$. Besides, the main effects of subscale length (L; $F_{(3, 4737)} = 18133.95$, $p < .001$, $\eta_G^2 = .91$), difficulty distribution (D; $F_{(2, 4737)} = 677.37$, $p < .001$, $\eta_G^2 = .21$), and inter-subscale correlations (R; $F_{(3, 4737)} = 390.96$, $p < .001$, $\eta_G^2 = .18$) were also significant.

To follow up significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. With regards to the effect of diagnostic method, results indicated that both AIRT and MIRT yielded average CCR of .827 which was significantly higher ($p < .001$) than that of ACDM ($M = .815$). With respect to subscale length, all

pairwise comparisons were statistically significant ($p < .001$); CCR increased with subscale length with average CCR of .754, .808, .854, and .877 for 5, 10, 20, and 30 items, respectively. Likewise, all pairwise comparisons were statistically significant ($p < .001$) with respect to difficulty distribution; CCR increased with difficulty distribution changing from HM ($M = .815$) to HV ($M = .820$), and to BL ($M = .833$). All pairwise comparisons of inter-subscore correlations were also significant; CCR increased with inter-subscale correlations with average CCR of .815, .819, .826, and .832 for inter-subscore correlations of .30, .50, .70, and .80, respectively.

Table 4.6

Four-Way Mixed ANOVA Results of Correct Classification Rate

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.07	6744.81	<.001	.12
DM x L	3.22	37.80	<.001	<.001
DM x D	2.15	1626.09	<.001	.06
DM x R	3.22	3.23	<.001	<.001
DM x L x D	6.44	6.07	<.001	<.001
DM x L x R	9.66	3.91	<.001	<.001
DM x D x R	6.44	0.53	.80	<.001
DM x L x D x R	19.32	0.17	>.99	<.001
Error	5085.44			
Between-subjects				
L	3	18133.95	<.001	.91
D	2	677.37	<.001	.21
R	3	390.96	<.001	.18
L x D	6	5.27	<.001	.01
L x R	9	22.27	<.001	.04
D x R	6	0.73	.63	<.001
L x D x R	18	0.04	>.99	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Interaction Effects. In addition to these main effects, the two-way interaction effect of diagnostic methods and item difficulty was also significant, $F_{(2.15,\ 5085.44)} = 1626.09$, $p < .001$, $\eta_G^2 = .061$ (Table 4.6). This interaction effect was first followed up by a simple main effect approach where the effect of the diagnostic model was examined at fixed levels of item difficulty using one-way repeated measures ANOVA. The simple main effects of the diagnostic model did not

meet the effect size criteria ($\eta_G^2 > .06$) at any levels of item difficulty although there existed differences in effect sizes between HM ($\eta_G^2 = .043$) versus the other levels ($\eta_G^2 = .004$ for HV and $\eta_G^2 = .002$ for BL) (Table 4.7). Accordingly, marginal means of CCR were relatively comparable between diagnostic methods at each level of item difficulty distribution (Table 4.8). These main and interaction effects of manipulated factors are illustrated in Figure 4.2 as well.

Table 4.7

One-Way Repeated Measures ANOVA Results of Correct Classification Rate at Fixed Levels of Item Difficulty Distribution

D	Factors	*df*n	*df*d	*F*	*p*	η_G^2
BL	Method	1.15	1833.59	993.72	<.001	.002
HV	Method	1.10	1746.02	1024.30	<.001	.004
HM	Method	1.05	1682.65	4610.17	<.001	.043

Note. D = item difficulty distribution, *df*n = *df* for numerator (i.e., *df* for mean square explained by the different groups [$MS_{between}$]), *df*d = *df* for denominator (i.e., *df* for mean square that is due to chance [MS_{within}]), η_G^2 = generalized eta squared, BL = baseline, HV = high variability, HM = high mean.

Table 4.8

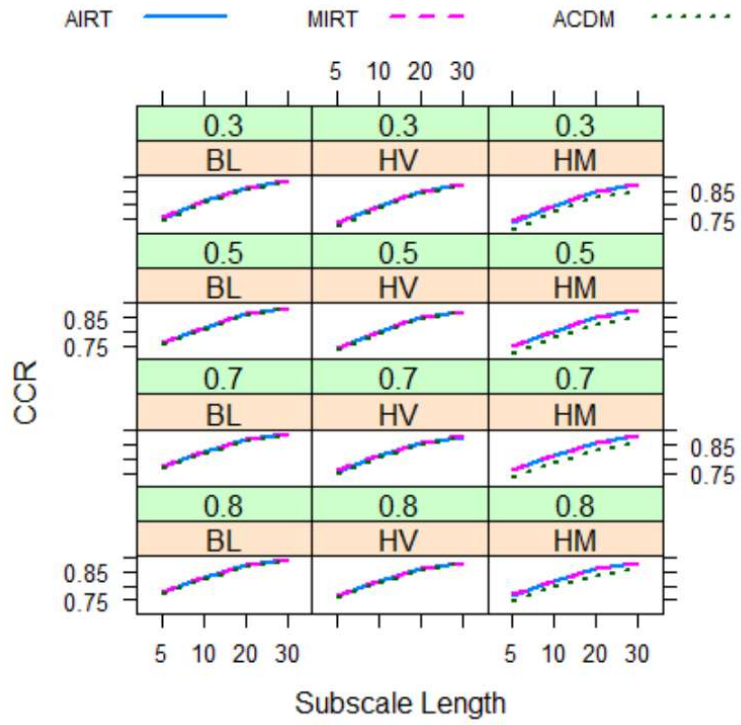
Marginal Means of Correct Classification Rate for Each Diagnostic Method by Item Difficulty Distribution

D	AIRT	MIRT	ACDM
BL	.84	.84	.83
HV	.82	.82	.82
HM	.82	.82	.80

Note. D = item difficulty distribution, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Figure 4.2

Average Correct Classification Rate by Simulation Conditions



Note. CCR = correct classification rate, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Kappa

Descriptive. Table 4.9 contains the average Kappa by each level of manipulated factors. On average, IRT-based methods yielded comparable levels of kappa (M = .65 for AIRT and MIRT) that was higher than that of ACDM (M = .63) across conditions. Kappa also increased with subscale length and inter-subscale correlations for all estimation methods. Besides, kappa increased as item difficulty distribution became more aligned with examinee ability distribution with the highest value at BL followed by HV and HM. As was the case with CCR, the difference between the kappa of IRT-based methods versus ACDM was relatively more pronounced at

HM. These patterns of relationships were consistent with findings for CCR. However, kappa was substantially lower than CCR across conditions. This is because kappa corrects the evaluation bias by considering the correct classification by a random guess, whereas CCR does not.

Table 4.9

Average Kappa by Manipulated Factors

MF	Level	N	AIRT	MIRT	ACDM
L	5	1199	.52	.52	.49
	10	1197	.62	.62	.60
	20	1192	.72	.72	.69
	30	1197	.76	.76	.74
D	BL	1599	.67	.67	.66
	HV	1587	.65	.65	.63
	HM	1599	.65	.65	.60
R	.30	1196	.64	.64	.61
	.50	1197	.64	.65	.62
	.70	1197	.66	.66	.64
	.80	1195	.67	.67	.65
All		4785	.65	.65	.63

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Main Effects. Table 4.10 contains mixed ANOVA results of kappa with Greenhouse-Geisser correction. Based on *p*-value ($p < .05$) and effect size criteria ($\eta_G^2 > .06$), the main effect of diagnostic methods (DM) was significant $F_{(1.07, 5075.58)} = 6918.40$, $p < .001$, $\eta_G^2 = .12$. Besides, the main effects of subscale length (L; $F_{(3, 4737)} = 18088.48$, $p < .001$, $\eta_G^2 = .91$), item difficulty

distribution (D; $F_{(2, 4737)} = 679.62$, $p < .001$, $\eta_G^2 = .21$), and inter-subscale correlations (R; $F_{(3, 4737)} = 391.58$, $p < .001$, $\eta_G^2 = .18$) were also significant.

To follow up on significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. Concerning the diagnostic method effect, results indicated that both AIRT and MIRT yielded an average kappa of .65 which was (statistically) significantly larger ($p < .001$) than that of ACDM ($M = .63$). Concerning subscale length, all pairwise comparisons were statistically significant ($p < .001$); kappa increased with subscale length with average kappa of .51, .62 .71, and .75 for 5, 10, 20, and 30 items, respectively. Likewise, all pairwise comparisons were statistically significant ($p < .001$) with respect to difficulty distribution; kappa increased with item difficulty distribution changing from HM ($M = .63$) to HV ($M = .64$), and to BL ($M = .67$). All pairwise comparisons of inter-subscore correlations were also significant; kappa increased with inter-subscale correlations with average CCR of .63, .64, .65, and .67 for inter-subscore correlations of .30, .50, .70, and .80, respectively. Generally, these results were parallel to those of CCR.

Table 4.10

Four-Way Mixed ANOVA Results of Kappa

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.07	6918.40	<.001	.12
DM x L	3.21	40.02	<.001	<.001
DM x D	2.14	1673.79	<.001	.06
DM x R	3.21	3.41	.02	<.001
DM x L x D	6.43	6.50	<.001	<.001
DM x L x R	9.64	4.01	<.001	<.001
DM x D x R	6.43	0.51	.81	<.001
DM x L x D x R	19.29	0.17	>.99	<.001
Error	5075.58			
Between-subjects				
L	3	18088.48	<.001	.91
D	2	679.62	<.001	.21
R	3	391.58	<.001	.18
L x D	6	5.13	<.001	.01
L x R	9	22.26	<.001	.04
D x R	6	0.74	.62	<.001
L x D x R	18	0.04	>.99	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Interaction Effects. The two-way interaction effect of diagnostic method and item difficulty distribution was also significant, $F_{(2.14,\ 5075.58)} = 1673.79$, $p < .001$, $\eta_G^2 = .06$ (Table 4.10). This interaction effect was followed up by a simple main effect approach where the effect of the diagnostic method was examined at fixed levels of item difficulty distribution using one-way repeated measures ANOVA. The simple main effects of the diagnostic method did not meet

the effect size criteria ($\eta_G^2 > .06$) at any level of item difficulty distribution although there existed differences in effect sizes between HM ($\eta_G^2 = .043$) versus the other levels ($\eta_G^2 = .004$ for HV and $\eta_G^2 = .002$ for BL; Table 4.11). Consistent with these results, marginal means of kappa were relatively comparable between diagnostic methods at each level of item difficulty distribution (Table 4.12). These main and interaction effects of manipulated factors are illustrated in Figure 4.3 as well.

Table 4.11

One-Way Repeated Measures ANOVA Results of Kappa at Fixed Levels of Item Difficulty Distribution

D	Factors	*df*n	*df*d	*F*	*p*	η_G^2
BL	Method	1.14	1828.38	1019.02	<.001	.002
HV	Method	1.10	1742.67	1048.11	<.001	.004
HM	Method	1.05	1679.38	4714.26	<.001	.043

Note. D = item difficulty distribution, *df*n = *df* for numerator (i.e., *df* for mean square explained by the different groups [$MS_{between}$]), *df*d = *df* for denominator (i.e., *df* for mean square that is due to chance [MS_{within}]), η_G^2 = generalized eta squared, BL = baseline, HV = high variability, HM = high mean.

Table 4.12

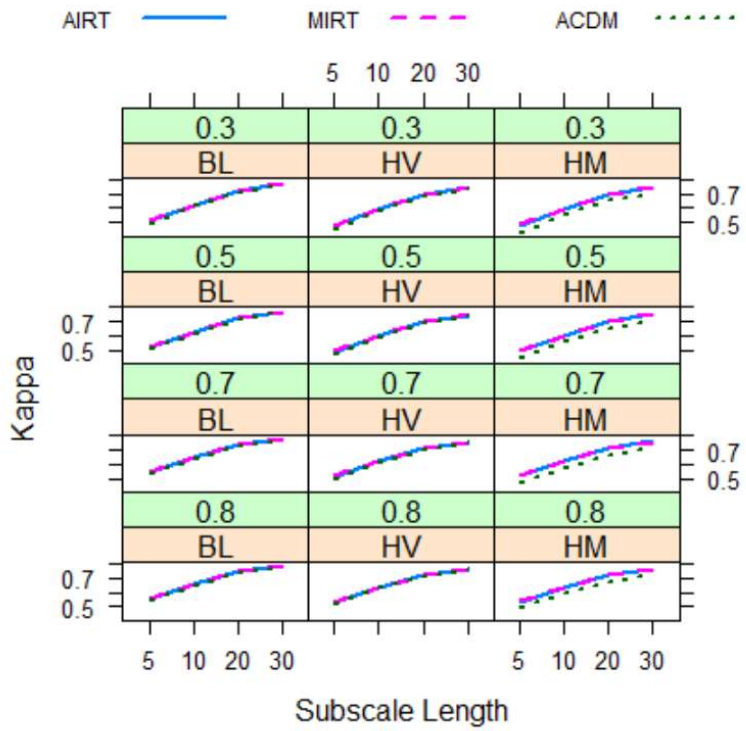
Marginal Means of Kappa for Each Diagnostic Method by Item Difficulty Distribution

D	AIRT	MIRT	ACDM
BL	.67	.67	.66
HV	.65	.65	.63
HM	.65	.65	.60

Note. D = item difficulty distribution, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Figure 4.3

Average Kappa by Simulation Condition



Note. AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Sensitivity

Descriptive. Table 4.13 contains the average sensitivity by each level of manipulated factors. On average, IRT-based methods yielded comparable levels of sensitivity (M = .83 for AIRT and MIRT) that was higher than that of ACDM (M = .78). Sensitivity also increased with subscale length and inter-subscale correlations for all estimation methods. Besides, it increased as item difficulty distribution became more aligned with examinee ability distribution with the highest value at BL followed by HV and HM. As in CCR and kappa, sensitivity tended to be comparable across diagnostic methods at BL and HV. However, there was a substantial

difference between IRT-based methods versus ACDM at HM (M = .83, .82, and .68 for AIRT, MIRT, and ACDM, respectively).

Table 4.13

Average Sensitivity by Manipulated Factors

MF	Level	N	AIRT	MIRT	ACDM
L	5	1199	.76	.76	.70
	10	1197	.81	.81	.76
	20	1192	.86	.86	.81
	30	1197	.88	.88	.84
D	BL	1599	.84	.84	.83
	HV	1587	.82	.82	.82
	HM	1599	.83	.82	.68
R	.30	1196	.82	.82	.77
	.50	1197	.83	.82	.77
	.70	1197	.83	.83	.78
	.80	1195	.84	.84	.79
All		4785	.83	.83	.78

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Main Effects. Table 4.14 contains mixed ANOVA results of sensitivity with Greenhouse-Geisser correction. Based on p-value ($p < .05$) and effect size criteria ($\eta_G^2 > .06$), the main effects of diagnostic methods (DM; $F_{(1.02, 4842.59)} = 4041.127$, $p < .001$, $\eta_G^2 = .31$), subscale length (L; $F_{(3, 4737)} = 6010.82$, $p < .001$, $\eta_G^2 = .64$), and item difficulty distribution (D; $F_{(2, 4737)} = 2238.22$, $p < .001$, $\eta_G^2 = .31$), were significant.

To follow up on significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. Concerning the diagnostic method effect,

results indicated that both AIRT and MIRT yielded average sensitivity of .83 which was significantly larger ($p < .001$) than that of ACDM ($M = .78$). With respect to subscale length, all pairwise comparisons were statistically significant ($p < .001$); sensitivity increased with subscale length with average of .74, .80 .84, and .87 for 5, 10, 20, and 30 items, respectively. Likewise, all pairwise comparisons were statistically significant ($p < .001$) with respect to item difficulty distribution; sensitivity increased with item difficulty distribution changing from HM ($M = .78$) to HV ($M = .82$), and to BL ($M = .83$).

Table 4.14

Four-Way Mixed ANOVA Results of Sensitivity

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.02	4041.13	<.001	.31
DM x L	3.07	32.84	<.001	.01
DM x D	2.04	3373.06	<.001	.43
DM x R	3.07	2.14	.092	<.001
DM x L x D	6.13	28.59	<.001	.02
DM x L x R	9.20	0.52	.866	<.001
DM x D x R	6.13	2.06	.054	<.001
DM x L x D x R	18.40	0.14	>.999	<.001
Error	4842.59			
Between-subjects				
L	3	6010.82	<.001	.64
D	2	2238.22	<.001	.31
R	3	131.25	<.001	.04
L x D	6	9.49	<.001	.01
L x R	9	8.53	<.001	.01
D x R	6	1.21	.299	<.001
L x D x R	18	0.10	>.999	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Interaction Effects. The two-way interaction effect of diagnostic method and item difficulty distribution was significant, $F_{(2.04, 4842.59)} = 3373.06$, $p < .001$, $\eta_G^2 = .43$ (Table 4.14). This interaction effect was followed up by a simple main effect approach where the effect of the diagnostic method was examined at fixed levels of item difficulty distribution using one-way repeated measures ANOVA. Simple main effects of M were significant only at HM, $F_{(1.04, 1654.17)}$

= 8663.62, $p < .001$, $\eta_G^2 = .515$ (Table 4.15). The significant simple main effect of the diagnostic method was followed up by simple comparisons where pairwise *t*-tests of the diagnostic method were performed at HM with Bonferroni adjustment. Results indicated that under HM conditions, all pairwise comparisons of diagnostic methods were statistically significant. Marginal means of sensitivity were highest for AIRT ($M = .83$) followed by MIRT ($M = .82$) and lowest for ACDM ($M = .68$) at HM, whereas differences between diagnostic methods were negligible at BL and HV (Table 4.16). These main and interaction effects of manipulated factors are illustrated in Figure 4.4 as well.

Table 4.15

One-Way Repeated Measures ANOVA Results of Sensitivity at Fixed Levels of Difficulty Distribution

D	Factors	*df*n	*df*d	*F*	*p*	η_G^2
BL	Method	1.02	1626.45	18.26	<.001	.002
HV	Method	1.03	1629.77	6.42	0.03	.001
HM	Method	1.04	1654.17	8663.62	<.001	.515

Note. η_G^2 = generalized eta squared, *df*n = *df* for numerator (i.e., *df* for mean square explained by the different groups [$MS_{between}$]), *df*d = *df* for denominator (i.e., *df* for mean square that is due to chance [MS_{within}]), D = item difficulty distribution, BL = baseline, HV = high variability, HM = high mean.

Table 4.16

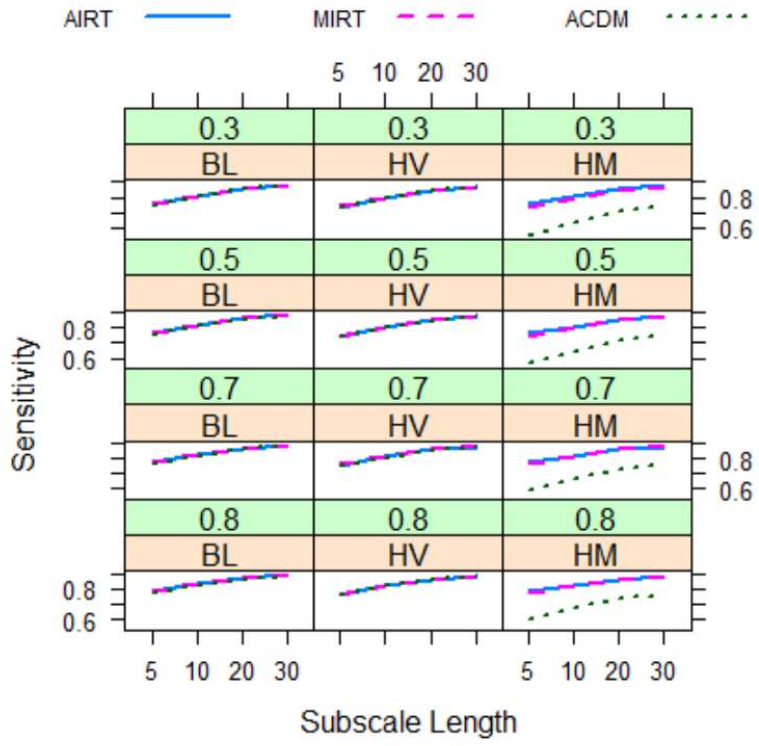
Marginal Means of Sensitivity for Each Diagnostic Method by Item Difficulty Distribution

D	AIRT	MIRT	ACDM
BL	.836	.836	.831
HV	.823	.824	.820
HM	.830	.823	.680

Note. D = item difficulty distribution, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Figure 4.4

Average Sensitivity by Simulation Conditions



Note. AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Specificity

Descriptive. Table 4.17 contains average specificity by each level of manipulated factors. On average, IRT-based methods yielded comparable levels of specificity that was lower than that of ACDM. Specificity also increased with subscale length and inter-subscale correlations for all estimation methods. With respect to item difficulty distribution, unlike other agreement measures, specificity was on average highest at HM followed by BL and HV. This was because ACDM yielded substantially higher specificity at HM compared to other conditions. IRT-based methods, in contrast, yielded specificity that was more consistent across levels of item difficulty distribution although it was highest at BL followed by HV and HM, as was the case with other

agreement measures. As a result, specificity tended to be comparable across diagnostic methods at BL and HV but was substantially higher for ACDM at HM.

Table 4.17

Average Specificity by Manipulated Factors

MF	Level	N	AIRT	MIRT	ACDM
L	5	1199	.75	.76	.79
	10	1197	.81	.81	.84
	20	1192	.86	.86	.88
	30	1197	.88	.88	.90
D	BL	1599	.83	.83	.83
	HV	1587	.82	.82	.81
	HM	1599	.82	.82	.92
R	.30	1196	.82	.82	.85
	.50	1197	.82	.82	.85
	.70	1197	.83	.83	.86
	.80	1195	.83	.84	.86
All		4785	.82	.83	.85

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Main Effects. Table 4.18 contains mixed ANOVA results of specificity with Greenhouse-Geisser correction. Based on p-value ($p < .05$) and effect size criteria ($\eta_G^2 > .06$), the main effects of diagnostic methods (DM; $F_{(1.03, 4892.94)} = 1595.96$, $p < .001$, $\eta_G^2 = .16$), subscale length (L; $F_{(3, 4737)} = 7418.34$, $p < .001$, $\eta_G^2 = .68$), and item difficulty distribution (D; $F_{(2, 4737)} = 1091.92$, $p < .001$, $\eta_G^2 = .17$), were significant.

To follow up significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. With respect to the diagnostic method effect,

all pairwise comparisons between AIRT (M = .824), MIRT (M = .827), and ACDM M (= .854) were statistically significant although the comparison between AIRT and MIRT was only marginally significant (p = .048 with Bonferroni adjustment). With respect to subscale length, all pairwise comparisons were statistically significant (p < .001); specificity increased with subscale length with average of .77, .82 .87, and .89 for 5, 10, 20, and 30 items, respectively. Likewise, all pairwise comparisons were statistically significant (p < .001) with respect to item difficulty distribution; specificity was highest at HM (M = .85) followed by BL (M = .83) and HV (M = .82).

Table 4.18

Four-Way Mixed ANOVA Results of Specificity

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.03	1595.96	<.001	.16
DM x L	3.10	23.06	<.001	.01
DM x D	2.07	2503.37	<.001	.37
DM x R	3.10	1.15	.33	<.001
DM x L x D	6.20	47.37	<.001	.03
DM x L x R	9.30	0.13	>.999	<.001
DM x D x R	6.20	2.52	.02	<.001
DM x L x D x R	18.59	0.18	>.999	<.001
Error	4892.94			
Between-subjects				
L	3	7418.34	<.001	.68
D	2	1091.92	<.001	.17
R	3	159.96	<.001	.04
L x D	6	18.82	<.001	.01
L x R	9	7.93	<.001	.01
D x R	6	1.08	.38	<.001
L x D x R	18	0.15	>.999	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Interaction Effects. The two-way interaction effect of diagnostic method and item difficulty distribution was also significant, $F_{(2.047, 4892.94)} = 2503.37$, $p < .001$, $\eta_G^2 = .37$ (Table 4.18). This interaction effect was followed up by a simple main effect approach where the effect of the diagnostic method was examined at fixed levels of item difficulty using one-way repeated measures ANOVA. Simple main effects of M were significant only at HM, $F_{(1.08, 1728.36)} =$

8234.97, $p < .001$, $\eta_G^2 = .483$ (Table 4.19). The significant simple main effect of the diagnostic method was followed up by simple comparisons where pairwise *t*-tests of the diagnostic method were performed at HM with Bonferroni adjustment. Results indicated that under HM conditions, all pairwise comparisons of diagnostic methods were statistically significant. Marginal means of specificity were highest for ACDM ($M = .92$) followed by MIRT ($M = .83$) and lowest for AIRT ($M = .812$) at HM, whereas differences between diagnostic methods were negligible at BL and HV (Table 4.20). These main and interaction effects of manipulated factors are illustrated in Figure 4.5 as well.

Table 4.19

One-Way Repeated Measures ANOVA Results of Specificity at Fixed Levels of Item Difficulty Distribution

D	Factors	*df*n	*Df*d	*F*	*p*	η_G^2
BL	Method	1.02	1628.40	18.12	<.001	.002
HV	Method	1.03	1628.54	38.76	<.001	.005
HM	Method	1.08	1728.36	8234.97	<.001	.483

Note. D = item difficulty distribution, *df*n = *df* for numerator (i.e., *df* for mean square explained by the different groups [$MS_{between}$]), *df*d = *df* for denominator (i.e., *df* for mean square that is due to chance [MS_{within}]), η_G^2 = generalized eta squared, BL = baseline, HV = high variability, HM = high mean.

Table 4.20

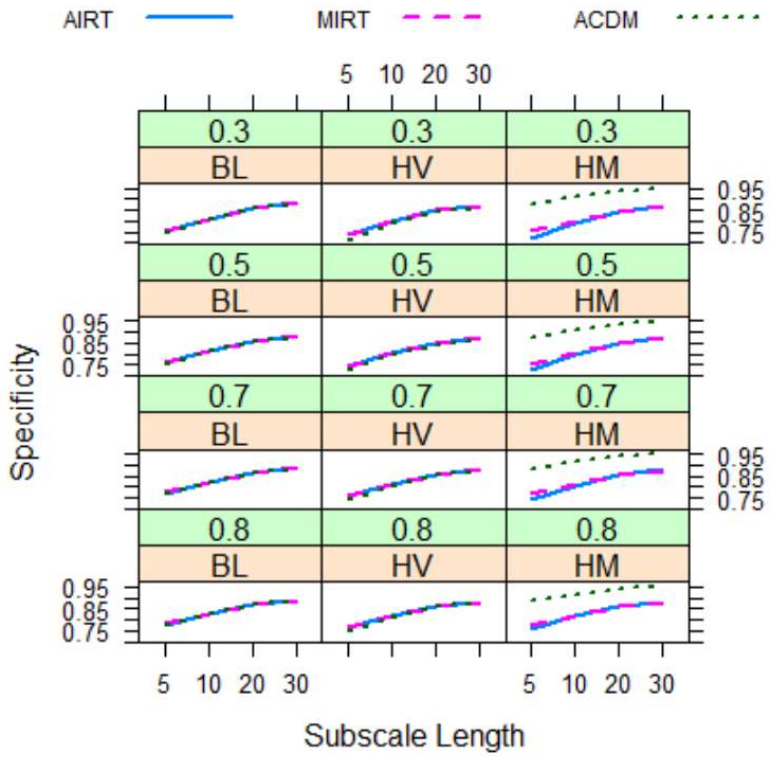
Marginal Means of Specificity for Each Diagnostic Method by Item Difficulty Distribution

D	AIRT	MIRT	ACDM
BL	.834	.834	.829
HV	.821	.822	.812
HM	.816	.824	.921

Note. D = item difficulty distribution, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Figure 4.5

Average Specificity by Simulation Conditions



Note. AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, BL = baseline, HV = high variability, HM = high mean.

Results for Research Question 2

Research question 2 concerns the extent to which subscores are distinct from each other as well as from overall score estimates in terms of discrete mastery categories and continuous proficiency scores. To address this question, inter-subscore correlations of each method and average correlations between subscores and overall scores were computed using both discrete mastery categories and continuous proficiency scores. Note that given the symmetry of two subscores in the current study design, correlations between each subscore and overall scores were averaged over two subdomains. Correlations based on discrete mastery categories were obtained

by tetrachoric correlations, whereas those based on continuous proficiency scores were obtained by Pearson correlations.

Inter-Subscore Correlations

Average intercorrelations of mastery categories were comparable between diagnostic methods with slightly higher tetrachoric correlations for AIRT and ACDM (M = .70 for both methods) versus MIRT (M = .68) (Table 4.21). In contrast, Pearson correlations among proficiency scores were comparable for AIRT (M = .70) and MIRT (M = .69) but lower for ACDM (M = .60), which is in part attributable to the fact that proficiency scores of ACDM have more restricted variability ranging from 0 to 1. When the PPM of ACDM was transformed to the logit scale using log (PPM / (1 - PPM)), the average inter-subscale correlation of ACDM increased to .64, but it was still slightly lower than those of IRT-based methods.

In addition to the diagnostic method, other factors also influenced inter-subscore correlations. As subscales get longer, intercorrelations of both discrete and continuous subscores decreased and the magnitude of decrease was greater for ACDM than for IRT-based methods. As expected, intercorrelations of subscores also increased with true inter-subscale correlations. In contrast, intercorrelations of subscores tended to be comparable across varying levels of item difficulty distribution although they were slightly lower at BL compared to HM and HV. In general, three manipulated factors (i.e., L, D, and R) exerted similar impact on each diagnostic method and the pattern of relationship was consistent regardless of type of subscores (i.e., discrete or continuous subscores). Inter-subscore correlations by simulation conditions are illustrated in Figure 4.6 for proficiency scores (i.e., Pearson correlations).

Correlations between Overall Scores and Subscores

Average correlations between overall scores and subscores ranged from .18 to 1.00 for mastery categories and from .28 to 1.00 for proficiency scores. On average, each diagnostic

method yielded comparable correlations with overall score estimates regarding discrete mastery categories (M = .91, .90, and .91 for AIRT, MIRT, and ACDM; Table 4.21). However, as was the case with intercorrelations of subscores, there was a difference between ACDM (M = .81 for PPM and M = .88 for logit transformed PPM) versus IRT-based methods (M = .91) with respect to correlations using proficiency scores.

Unlike the results of inter-subscore correlations, average correlations between overall scores and subscores only slightly decreased as subscale length increased although such a pattern was more noticeable for Pearson correlations and ACDM. In addition, as was the case with intercorrelations of subscores, correlations between overall scores and subscores were parallel across different item difficulty distribution and increased with true inter-subscore correlations. Note that even under optimal conditions (i.e., L = 30, D = BL, and R = .30), average correlations between overall scores and subscores were at least as high as .80 for all methods regardless of the type of subscores. The only exception was the Pearson correlation between overall scores and proficiency scores of ACDM, for which the minimum correlation was .69 based on original proficiency scores and .75 for logit transformed subscores. Given that these correlations inform the extent to which subscores are distinctive, they shed some light on the results related to the incremental criterion-related validity evidence of subscores, which is presented in the subsequent section.

As was the case with inter-subscore correlations, three manipulated factors (i.e., L, D, and R) in general exerted similar impact on each diagnostic method and the pattern of relationship was consistent regardless of type of subscores (i.e., discrete or continuous subscores). Correlations between overall scores and subscores by simulation conditions are illustrated in Figure 4.7 for proficiency scores (i.e., Pearson correlations).

Table 4.21

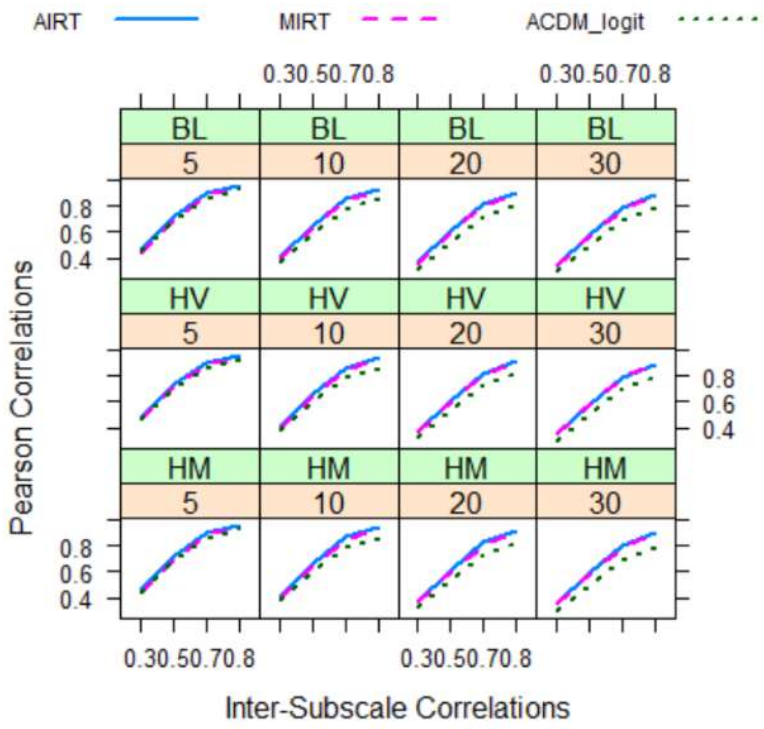
Inter-Subscore Correlations and Average Correlations between Overall Scores and Two Subscore by Manipulated Factors Using Discrete and Continuous Score Estimates

MF	Level	N	Tetrachoric correlations					
			AIRT	MIRT	ACDM	UIRT-AIRT	UIRT-MIRT	UIRT-ACDM
L	5	1199	.75	.73	.77	.91	.91	.91
	10	1197	.72	.69	.71	.91	.90	.91
	20	1192	.68	.65	.67	.91	.90	.91
	30	1197	.66	.64	.65	.90	.90	.90
D	BL	1599	.70	.67	.69	.91	.90	.91
	HV	1587	.70	.68	.71	.91	.91	.91
	HM	1599	.70	.68	.70	.91	.90	.90
R	.30	1196	.40	.38	.40	.81	.80	.81
	.50	1197	.64	.61	.64	.90	.89	.90
	.70	1197	.84	.82	.84	.96	.95	.95
	.80	1195	.92	.90	.91	.98	.97	.97
All		4785	.70	.68	.70	.91	.90	.91
IV	Level	N	Pearson correlations					
			AIRT	MIRT	ACDM	UIRT-AIRT	UIRT-MIRT	UIRT-ACDM
L	5	1199	.76	.74	.71 (.73)	.92	.92	.87 (.91)
	10	1197	.72	.70	.63 (.65)	.92	.91	.82 (.89)
	20	1192	.67	.66	.55 (.60)	.91	.91	.79 (.87)
	30	1197	.65	.64	.52 (.57)	.90	.90	.77 (.84)
D	BL	1599	.70	.68	.59 (.63)	.91	.91	.81 (.87)
	HV	1587	.70	.69	.61 (.64)	.91	.91	.81 (.88)
	HM	1599	.71	.69	.60 (.64)	.91	.91	.81 (.88)
R	.30	1196	.40	.39	.32 (.36)	.81	.81	.73 (.79)
	.50	1197	.64	.63	.53 (.58)	.90	.90	.80 (.87)
	.70	1197	.84	.83	.73 (.77)	.96	.95	.85 (.92)
	.80	1195	.92	.91	.82 (.84)	.98	.98	.86 (.93)
All		4785	.70	.69	.60 (.64)	.91	.91	.81 (.88)

Note. MF = manipulated factors, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean. UIRT-AIRT, UIRT-MIRT, and UIRT-ACDM refer to average correlations between overall scores and two subscores of AIRT, MIRT, and ACDM, respectively. Correlations using logit transformed proficiency scores of ACDM appear inside the parentheses.

Figure 4.6

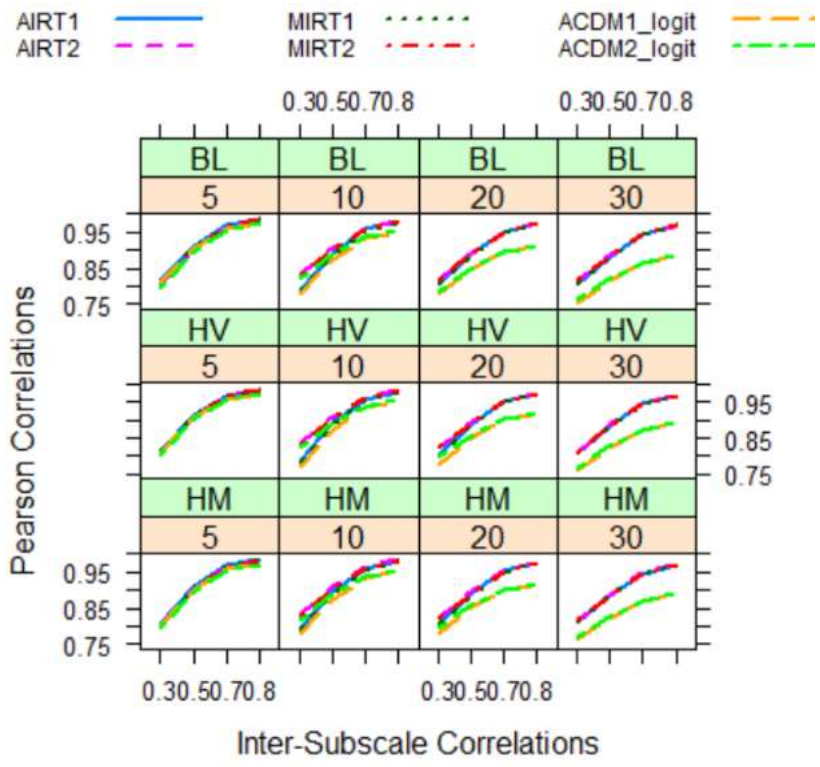
Inter-Subscore Pearson Correlations by Simulation Conditions



Note. AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM_logit = logit transformed proficiency scores of ACDM, BL = baseline, HV = high variability, HM = high mean

Figure 4.7

Pearson Correlations between Overall Scores and Subscores by Simulation Conditions



Note. AIRT1 and AIRT2 refer to subscores 1 and 2 estimated by augmented IRT, MIRT1 and MIRT2 refer to subscores 1 and 2 estimated by multidimensional IRT, ACDM1_logit and ACDM2_logit refer to logit transformed subscore 1 and 2 estimated by additive CDM, BL = baseline, HV = high variability, HM = high mean.

Results for Research Question 3

Research question 3 concerns the extent to which each diagnostic method provides added value or incremental criterion-related validity evidence over and above the UIRT. As stated earlier, M_1 refers to the restricted regression model with the overall score as the only predictor, and M_2 refers to the full regression model with two subscores plus overall scores as predictors. Note that unlike analyses for classification accuracy measures, ANOVA analyses of R^2 were performed on four estimation methods, three diagnostic methods plus UIRT, to allow for

comparisons between multidimensional versus unidimensional approaches. Therefore, the DVs for mixed ANOVA included R^2 of M_1 (for UIRT) and M_2 (for three diagnostic methods) for high and low criterion validity coefficient conditions (i.e., HCV and LCV). Subsequent sections are organized by the magnitude of criterion validity coefficients.

Results for High Criterion Validity Coefficients

Descriptive. Table 4.22 contains correlations of criterion variables with overall and subscores. In general, criterion correlations were highest for UIRT (M = .73) followed by AIRT and MIRT (M = .68 for both methods) and lowest for ACDM (M = .60 for original proficiency scores and M = .65 for logit transformed proficiency scores). Also, correlations of criterion variables with overall and subscores increased as subscales got longer, inter-subscale correlations increased, and item difficulty distribution changed from HM to HV and BL.

Table 4.23 contains R^2 of two regression models (i.e., M_1 and M_2) by manipulated factors for high criterion validity coefficient conditions. Three diagnostic methods yielded comparable R^2 for full regression models (M = .550, .550, and .548 for AIRT, MIRT, and ACDM) that was higher than that of the restricted regression model (M = .542). Accordingly, the change in R^2 between M_2 and M_1 ($\Delta R^2_{M_2-M_1}$) was on average comparable for three diagnostic methods although it was slightly higher for AIRT and MIRT (M = .008 for both methods) than for ACDM (M =.006). As expected, R^2 increased with subscale length and intercorrelations of subscores. Concerning difficulty distribution, R^2 increased as the item difficulty distribution became more aligned with examinee ability distribution, highest at BL followed by HV and lowest at HM. On contrary, $\Delta R^2_{M_2-M_1}$ tended to decrease as $R^2_{M_1}$ and $R^2_{M_2}$ increased. As such, $\Delta R^2_{M_2-M_1}$ slightly decreased with subscale length and substantially decreased with intercorrelations of subscores. On the other hand, no notable relationship was observed between $\Delta R^2_{M_2-M_1}$ and the item difficulty distribution.

Table 4.22

Correlations between Criterion Variables and Overall and Subscores by Manipulated Factors for High Criterion Validity Coefficient Condition

MF	Level	N	UIRT	AIRT	MIRT	ACDM
L	5	1199	.63	.60	.60	.56 (.59)
	10	1197	.72	.67	.67	.60 (.66)
	20	1192	.78	.72	.72	.62 (.68)
	30	1197	.80	.73	.73	.62 (.67)
D	BL	1599	.74	.69	.68	.61 (.66)
	HV	1587	.73	.68	.68	.60 (.65)
	HM	1599	.73	.68	.67	.59 (.65)
R	.30	1196	.69	.59	.58	.52 (.57)
	.50	1197	.73	.66	.66	.58 (.63)
	.70	1197	.75	.72	.72	.64 (.69)
	.80	1195	.76	.75	.75	.66 (.71)
All		4785	.73	.68	.68	.60 (.65)

Note. MF = manipulated factors, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean. Correlations using logit transformed ACDM estimates appear inside the parentheses.

Table 4.23

R^2 by Manipulated Factors for High Criterion Validity Coefficient Condition

MF	Level	N	M_1	M_2		
			UIRT	AIRT	MIRT	ACDM
L	5	1199	.397	.409 (.011)	.408 (.011)	.407 (.010)
	10	1197	.518	.527 (.009)	.527 (.009)	.525 (.007)
	20	1192	.610	.616 (.006)	.616 (.006)	.614 (.004)
	30	1197	.644	.649 (.005)	.649 (.005)	.647 (.003)
D	BL	1599	.555	.563 (.008)	.563 (.008)	.561 (.006)
	HV	1587	.537	.545 (.008)	.545 (.008)	.543 (.006)
	HM	1599	.535	.542 (.008)	.542 (.008)	.540 (.006)
R	.30	1196	.477	.499 (.023)	.500 (.023)	.495 (.018)
	.50	1197	.533	.539 (.006)	.540 (.006)	.538 (.005)
	.70	1197	.571	.573 (.002)	.573 (.002)	.572 (.001)
	.80	1195	.587	.588 (.001)	.588 (.001)	.588 (.000)
All		4785	.542	.550 (.008)	.550 (.008)	.548 (.006)

Note. MF = manipulated factors, M_1 = restricted regression model with overall scores as the only predictor, M_2 = full regression model with two subscores plus overall scores as predictors, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM. L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean. Numbers inside parentheses indicate the difference in adjusted R^2 between full and restricted models for the given diagnostic method.

Main Effects. Table 4.24 contains four-way mixed ANOVA results of R^2 for HCV with Greenhouse-Geisser correction. The main effect of diagnostic methods (DM) was statistically significant but did not meet the effect size criteria, $F_{(1.06, 5032.20)} = 2912.32$, $p < .001$, $\eta_G^2 = .02$. However, such insignificance may be attributable to the comparable R^2 between the three diagnostic methods although all diagnostic methods yielded R^2 that was (statistically) significantly higher than that of UIRT (Table 4.23). The main effects of subscale length (L; $F_{(3, 4737)} = 21110.70$, $p < .001$, $\eta_G^2 = .93$), item difficulty distribution (D; $F_{(2, 4737)} = 279.27$, $p < .001$,

η_G^2= .10), and inter-subscale correlations (R; $F_{(3, 4737)}$ = 3229.92, p < .001, $\eta_G^{2\,2}$ = .67) were all significant. However, none of the interaction effects were significant.

To follow up on significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. Concerning subscale length, all pairwise comparisons were statistically significant (p < .001); R^2 of M_1 and M_2 increased with subscale length having marginal means of .405, .525, .614, and .647 for 5, 10, 20, and 30 items, respectively. With respect to item difficulty distribution, pairwise comparisons between BL and the other two levels were statistically significant (p < .001) but the comparison between HV and HM was not significant (p = .29); R^2 of M_1 and M_2 was higher at BL (M = .560) compared to HV (M = .543) and HM (M = .540). With respect to inter-subscale correlations, all pairwise comparisons were statistically significant (p < .001); R^2 of M_1 and M_2 increased with inter-subscale correlations with marginal means of .493, .537, .572, and .588 for inter-subscale correlations of .30, .50, .70, and .80, respectively. Besides, in supplemental analyses for the estimation method, all pairwise comparisons between three diagnostic methods (M = .550, .550, and .548 for AIRT, MIRT, and ACDM) versus UIRT (M = .542) were statistically significant, whereas all pairwise comparisons between three diagnostic methods were insignificant (p > .999).

Table 4.24

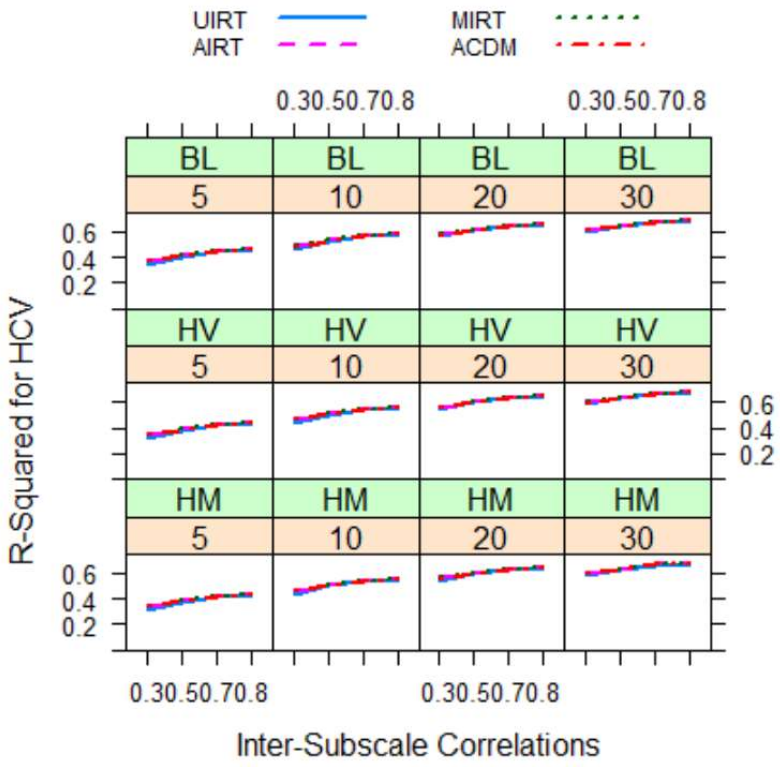
Four-Way Mixed ANOVA Results of R^2 for High Criterion Validity Coefficient Condition

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.06	2912.32	<.001	.02
DM x L	3.19	122.70	<.001	.00
DM x D	2.12	1.29	.28	<.001
DM x R	3.19	1178.17	<.001	.02
DM x L x D	6.37	1.03	.41	<.001
DM x L x R	9.56	45.39	<.001	<.001
DM x D x R	6.37	0.48	.84	<.001
DM x L x D x R	19.12	0.48	.97	<.001
Error	5032.20			
Between-subjects				
L	3	21110.70	<.001	.93
D	2	279.27	<.001	.10
R	3	3229.92	<.001	.67
L x D	6	8.75	<.001	.01
L x R	9	5.92	<.001	.01
D x R	6	0.08	>.999	<.001
L x D x R	18	0.05	>.999	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Figure 4.8

Average R^2 by Simulation Conditions for High Criterion Validity Coefficient Condition



Note. UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, HCV = high criterion validity, BL = baseline, HV = high variability, HM = high mean.

Results for Low Criterion Validity Coefficients

Descriptive. Table 4.25 contains correlations of criterion variables with overall and subscores. In general, criterion correlations were highest for UIRT (M = .50) followed by AIRT and MIRT (M = .46 for both methods) and lowest for ACDM (M = .41 for original proficiency scores and M = .44 for logit transformed proficiency scores). Also, correlations of criterion variables with overall and subscores increased as subscales got longer, inter-subscale correlations increased, and item difficulty distribution changed from HM to HV and BL.

Table 4.26 contains the R^2 of two regression models (i.e., M_1 and M_2) by manipulated factors for low criterion validity coefficient conditions. Three diagnostic methods yielded comparable R^2 for full regression models (M = .257, .257, and .256 for AIRT, MIRT, and ACDM) that was higher than that of the restricted regression model (M = .253). Accordingly, the change in R^2 between M_2 and M_1 ($\Delta R^2_{M_2-M_1}$) was on average comparable for three diagnostic methods although it was slightly higher for AIRT and MIRT (M = .003 for both methods) than for ACDM (M =.002). As expected, R^2 increased with subscale length and intercorrelations of subscores. Concerning difficulty distribution, R^2 increased as the item difficulty distribution became more aligned with examinee ability distribution, highest at BL followed by HV and lowest at HM. On contrary, $\Delta R^2_{M_2-M_1}$ tended to decrease as $R^2_{M_1}$ and $R^2_{M_2}$ increased. As such, $\Delta R^2_{M_2-M_1}$ slightly decreased with subscale length and substantially decreased with intercorrelations of subscores. On the other hand, no notable relationship was observed between $\Delta R^2_{M_2-M_1}$ and the item difficulty distribution. In general, these patterns of relationships were consistent between high and low criterion validity coefficient conditions. However, both $R^2_{M_2}$ and $\Delta R^2_{M_2-M_1}$ were lower at low criterion validity coefficient condition compared to high criterion validity coefficient condition in each cell.

Table 4.25

Correlations between Criterion Variables and Overall and Subscores by Manipulated Factors for Low Criterion Validity Coefficient Condition

MF	Level	N	UIRT	AIRT	MIRT	ACDM
L	5	1199	.43	.41	.41	.38 (.40)
	10	1197	.49	.46	.46	.41 (.45)
	20	1192	.53	.49	.49	.42 (.47)
	30	1197	.55	.50	.50	.42 (.46)
D	BL	1599	.51	.47	.47	.41 (.45)
	HV	1587	.50	.46	.46	.41 (.44)
	HM	1599	.50	.46	.46	.41 (.44)
R	.30	1196	.45	.39	.38	.34 (.37)
	.50	1197	.49	.45	.45	.40 (.43)
	.70	1197	.52	.50	.50	.44 (.48)
	.80	1195	.54	.52	.52	.46 (.50)
All		4785	.50	.46	.46	.41 (.44)

Note. MF = manipulated factors, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean.

Table 4.26

R^2 by Manipulated Factors for Low Criterion Validity Coefficient Condition

MF	Level	N	M_1	M_2		
			UIRT	AIRT	MIRT	ACDM
L	5	1199	.185	.189 (.005)	.189 (.005)	.189 (.004)
	10	1197	.243	.247 (.004)	.247 (.004)	.246 (.003)
	20	1192	.285	.288 (.003)	.288 (.003)	.287 (.002)
	30	1197	.301	.303 (.002)	.303 (.002)	.302 (.001)
D	BL	1599	.260	.263 (.003)	.263 (.003)	.262 (.002)
	HV	1587	.251	.255 (.003)	.255 (.003)	.254 (.003)
	HM	1599	.249	.253 (.003)	.253 (.003)	.252 (.002)
R	.30	1196	.207	.216 (.009)	.216 (.009)	.214 (.007)
	.50	1197	.244	.247 (.003)	.247 (.003)	.246 (.002)
	.70	1197	.275	.275 (.001)	.275 (.001)	.275 (.000)
	.80	1195	.289	.289 (.000)	.289 (.000)	.289 (.000)
All		4785	.253	.257 (.003)	.257 (.003)	.256 (.002)

Note. MF = manipulated factors, M_1 = restricted regression model with overall scores as the only predictor, M_2 = full regression model with two subscores plus overall scores as predictors, UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM. L = subscale length, D = item difficulty distribution, R = inter-subscore correlations, BL = baseline, HV = high variability, HM = high mean. Numbers inside parentheses indicate the difference in adjusted R^2 between full and restricted models for the given diagnostic method.

Main Effects. Table 4.27 contains four-way mixed ANOVA results of R^2 for LCV with Greenhouse-Geisser correction. The main effect of diagnostic methods (DM) was statistically significant but did not meet the effect size criteria, $F_{(1.21, 5730.22)} = 2188.39$, $p < .001$, $\eta_G^2 = .01$. As was the case for the high criterion validity coefficient condition, such insignificance may be attributable to the comparable R^2 between the three diagnostic methods although all diagnostic methods yielded R^2 that was (statistically) significantly higher than that of UIRT (Table 4.26). The main effects of subscale length (L; $F_{(3, 4737)} = 8800.79$, $p < .001$, $\eta_G^2 = .85$) and inter-subscale correlations (R; $F_{(3, 4737)} = 3816.98$, $p < .001$, $\eta_G^2 = .71$) were significant, whereas the main effect

of item difficulty distribution was only marginally significant in terms of effect size criteria, $F_{(2, 4737)} = 130.73$, $p < .001$, $\eta_G^2 = .05$. As in high criterion validity coefficient condition, none of the interaction effects were significant.

To follow up significant main effects, pairwise comparisons of the estimated marginal means were conducted with Bonferroni adjustment. With respect to subscale length, all pairwise comparisons were statistically significant ($p < .001$); R^2 of M_1 and M_2 increased with subscale length having marginal means of .188, .246, .287, and .302 for 5, 10, 20, and 30 items, respectively. With respect to item difficulty distribution, pairwise comparisons between BL and the other two levels were statistically significant ($p < .001$) but the comparison between HV and HM was not significant ($p = .09$); R^2 of M_1 and M_2 was higher at BL ($M = .262$) compared to HV ($M = .254$) and HM ($M = .252$). With respect to inter-subscale correlations, all pairwise comparisons were statistically significant ($p < .001$); R^2 of M_1 and M_2 increased with inter-subscale correlations with marginal means of .213, .246, .275, and .289 for inter-subscale correlations of .30, .50, .70, and .80, respectively. Besides, in supplemental analyses for the estimation method, pairwise comparisons between IRT-based methods ($M = .257$ for both AIRT and MIRT) versus UIRT ($M = .253$) were statistically significant ($p = .02$), whereas comparisons between ACDM and UIRT ($p = .20$) as well as comparisons three diagnostic methods were insignificant ($p > .999$).

Table 4.27

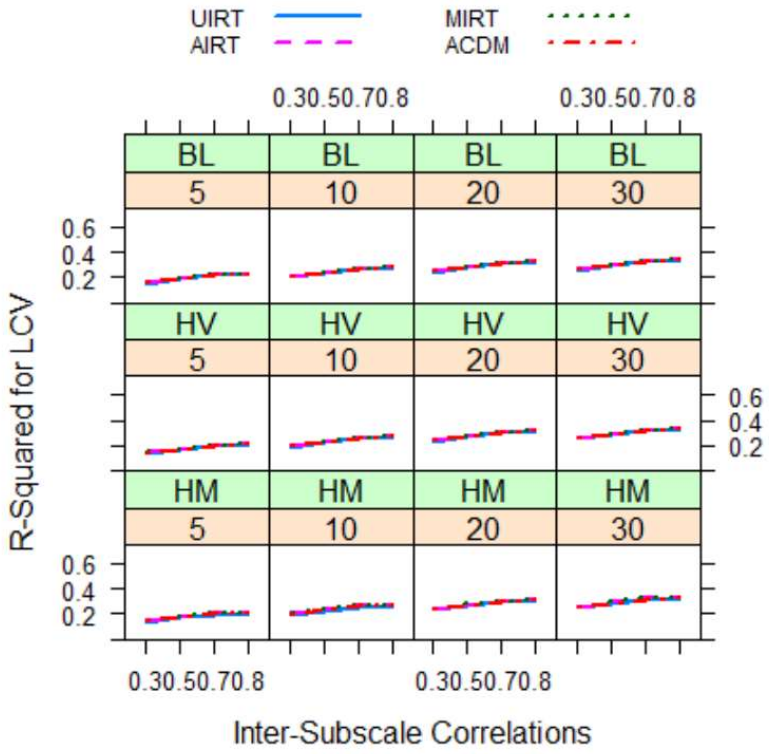
Four-Way Mixed ANOVA Results of R^2 for Low Criterion Validity Coefficient Condition

Source	*df*	*F*	*p*	η_G^2
Within-subjects				
DM	1.21	2188.39	<.001	.01
DM x L	3.63	84.97	<.001	<.001
DM x D	2.42	0.46	.67	<.001
DM x R	3.63	845.41	<.001	.01
DM x L x D	7.26	0.84	.56	<.001
DM x L x R	10.89	31.68	<.001	<.001
DM x D x R	7.26	0.23	.98	<.001
DM x L x D x R	21.77	0.42	.99	<.001
Error	5730.22			
Between-subjects				
L	3	8800.79	<.001	.85
D	2	130.73	<.001	.05
R	3	3816.98	<.001	.71
L x D	6	5.90	<.001	.01
L x R	9	2.12	.03	<.001
D x R	6	0.33	.92	<.001
L x D x R	18	0.07	>.999	<.001
Error	4737			

Note. η_G^2 = generalized eta squared, DM = diagnostic method, L = subscale length, D = item difficulty distribution, R = inter-subscore correlations.

Figure 4.9

Average R^2 by Simulation Conditions for Low Criterion Validity Coefficient Condition



Note. UIRT = unidimensional IRT, AIRT = augmented IRT, MIRT = multidimensional IRT, ACDM = additive CDM, LCV = low criterion validity, BL = baseline, HV = high variability, HM = high mean.

Chapter 5

Discussion

Despite the increasing demand for diagnostic information, researchers have often reported that observed subscores lacked adequate psychometric qualities such as reliability, distinctiveness, and validity. Among several statistical techniques to improve the quality of subscores estimates, AIRT and MIRT which utilize collateral information to improve subscore reliability have been endorsed as the most promising diagnostic tools. More recently, DCMs have also attracted increasing attention as a powerful diagnostic tool that can provide detailed diagnostic feedback with greater accuracy. Although DCMs may be a viable alternative to extant subscore estimation methods, there has been a dearth of research evaluating the psychometric quality of DCM estimates, especially in comparison with diagnostic subscores estimated from different frameworks. Therefore, in the present study, I aimed to compare DCMs with AIRT and MIRT under various conditions in terms of reliability, distinctiveness, and validity evidence of subscores.

Given that DCMs primarily provide discrete mastery classification in addition to the continuous probability estimates of mastery status, I utilized both discrete mastery categories and continuous proficiency scores for parallel comparisons between DCMs versus those IRT-based methods. First, I evaluated subscore reliability in terms of classification accuracy (research question 1) given that the reliability of discrete categories is evaluated in terms of classification accuracy or consistency. Besides, I evaluated subscore distinctiveness in terms of inter-subscore correlations and correlations between subscores and overall scores (research question 2) using both discrete mastery categories and continuous proficiency scores. Moreover, I examined the added value of subscores in terms of the incremental criterion-related validity evidence of subscores over and above overall scores (research question 3) using continuous proficiency

scores. In this section, results pertaining to each research question are discussed in order of three research questions. Discussion of results is followed by a summary of findings and limitations of the present study and future research directions. In addition, a conclusion along with some recommendations for choosing the appropriate subscore estimation methods and constructing diagnostic assessments are provided at the end.

Research Question 1: Classification Accuracy of Diagnostic Methods

In this section, findings related to the classification accuracy of subscores (research question 1) are discussed. This section is organized into two parts: comparative classification accuracy of diagnostic methods and the impact of various factors (i.e., L, D, and R) on the classification accuracy of subscores.

Comparative Classification Accuracy of Diagnostic Methods

For all measures of agreement, three diagnostic methods tended to yield comparable classification accuracy when average item difficulty was well aligned with average examinee ability and cut-scores. However, in the presence of a mismatch between average item difficulty and average examinee ability, ACDM resulted in classification accuracy that was significantly different from that of IRT-based methods, whereas AIRT and MIRT yielded comparable classification accuracy across all conditions. Besides, the pattern of difference between ACDM versus IRT-based methods varied depending on the type of classification measures. That is, for correct classification rates and kappa, the interaction effect of the diagnostic method and the item difficulty distribution was only marginally significant in terms of effect sizes ($p < .001$, $\eta_G^2 = .06$), and differences between ACDM and IRT-based methods were relatively small (i.e., average differences of 2.3 % and 4.6% for correct classification rates and kappa, respectively) even at HM. However, with respect to sensitivity and specificity, differences between ACDM and IRT-

based methods were substantial at HM; under this condition, ACDM yielded substantially lower sensitivity but substantially higher specificity when compared to IRT-based methods.

This was primarily because ACDM underestimated true mastery rates when tests were on average harder relative to examinee ability, whereas IRT-based methods accurately estimated true mastery rates only with slight deviations across all conditions. Given that true mastery rates were consistent across conditions, a substantial decrease in base mastery rates of ACDM led to a trade-off between sensitivity and specificity. That is, for ACDM, sensitivity – the proportion of true masters who are classified as masters (i.e., true positive or hit) – substantially decreased at HM, whereas specificity – the proportion of true non-masters who are classified as non-masters (i.e., true negative or correct rejection) – substantially increased at HM. Note that if tests were easier relative to examinee ability and the same cut-score of $\theta = 0$ was applied, the direction of such bias would be reversed. That is, ACDM would overestimate true mastery rates and thus yield higher sensitivity and lower specificity than would do IRT-based methods.

If DCMs are to provide objective classification or criterion-referenced interpretations, they should produce the same base mastery rates regardless of the test difficulty given that examinee ability distribution and cut-scores were held constant in this study. In other words, as in IRT, each examinee's mastery classification (or proficiency scores) should be determined solely based on ability estimates and should not change depending on the difficulty of the test examinees take. This property of test scores is referred to as item invariance. Because item invariance property did not hold for ACDM, each examinee's probability of attribute mastery decreased as tests became harder relative to the examinee's ability.

Such a trend is also reflected in the item parameters of ACDM; both intercepts (λ_0) and slope (λ_1) parameters of ACDM decreased as item difficulty increased (i.e., mean intercept = .34, .36, and .21 for BL, HV, and HM conditions, respectively; mean slopes = .33, .29, and .30

for BL, HV, and HM conditions, respectively). This means that on average, the probability that examinees who did not master the given attributes get the item correct is lower if they take harder tests (i.e., smaller intercepts). Likewise, on average, an increase in the probability of the correct item response brought about by mastery of each attribute for the item is lower if examinees take harder tests (i.e., smaller slopes).

These results are consistent with previous findings (e.g., Bradshaw & Madison, 2016; de la Torre & Lee, 2010) that the invariance property of DCMs does not hold under imperfect model-data fit. For example, using simulated data, Bradshaw and Madison (2016) examined the item-invariance property of DCMs, the extent to which examinee classifications remain the same regardless of the test difficulty, under model-data fit and model-data misfit conditions. To examine the item invariance property, they manipulated the test difficulty at three levels randomly drawing the difficulty parameters from $N(-1, 0.25)$, $N(0, 0.25)$, and $N(1, 0.25)$ distributions for the easy, moderate, and hard tests, respectively. The other model generating parameters such as examinee ability distribution (i.e., a multivariate normal distribution with a mean of zero and all pairwise ability correlations of 0.50) and discrimination parameters (i.e., all highly discriminating) were held constant across conditions. When data were generated and estimated by LCDM (i.e., model-data fit condition), they found negligible differences in correct classification rates and invariance classification rates (i.e., ICR; the proportion of examinees who received the same classification regardless of whether the classification is correct) between different test difficulty conditions with other things being held constant.

In contrast, when data were generated using M2PL but estimated by LCDM (i.e., model-data misfit condition), ICR values were substantially lower than those from the model-data fit conditions. Authors also found a clear trend that as the test difficulty increased, the estimated base rate of mastery decreased, even though the sample did not change. The estimated base rates

of mastery were 60.2%, 49.6%, and 38.3% for easy, moderate, and hard tests, respectively. These results demonstrated that classification accuracy and classification consistency depended on the model-data fit. Bradshaw and Madison (2016) further noted that model-data misfits caused by misspecifications in the Q-matrix are also likely to result in a substantial decrease in classification accuracy as shown by Rupp and Templin (2008a).

As shown earlier, ACDM yielded the poorest relative fit (i.e., the highest AIC and BIC) among four estimation methods given that data were generated based on IRT models in the present study. Therefore, it is not surprising that ACDM resulted in base mastery rates that substantially deviated from true mastery rates and those of IRT-based methods. Therefore, these results indicate that some additional procedures may be necessary when applying DCMs to tests constructed based on other psychometric frameworks (Chin, 2011). For example, linking/equating (e.g., Roussos et al., 2005; Xin & Zhang, 2015; X. Xu & von Davier, 2008) can be considered to infer criterion-referenced interpretations across different test forms and populations. Additionally, standard-setting procedures (e.g., Henson & Templin, 2008) deriving a series of cut scores that are typically pre-specified relative to population means can be considered to make fair comparisons of skill mastery estimates across different populations.

Impact of Various Factors on Classification Accuracy

The main effects of subscale length, item difficulty distribution, and inter-subscale correlations were significant in this study. First, classification accuracy increased with subscale length, which was consistent with findings from previous simulation studies (e.g., Lathrop & Cheng, 2013; Wan et al., 2007; Wyse, 2011). This is because longer tests are likely to provide more information about the examinee's ability, thus leading to greater score variability and allowing for a more accurate and stable estimation of ability. In line with it, Sinharay et al. (2019) noted that subscores based on a limited number of items don’t provide adequate

information to differentiate the examinee's ability on multiple subdomains although the cognitive theory of response processes indicates multiple subdomains underlying test data. Similarly, the classification accuracy of three diagnostic methods increased with intercorrelations of subscores. This is expected given that the composite is more likely to become consistent and stable as subscores become more correlated (Bulut et al., 2016; Davison et al., 2015). Subscale length and inter-subscale correlations did not interact with diagnostic methods such that their impact on classification accuracy was consistent across three diagnostic methods.

On the other hand, correct classification rates, kappa, and sensitivity, classification accuracy decreased when there was larger variability in item difficulty and when there was the mismatch between average item difficulty and average examinee ability compared to when average item difficulty matched with average examinee ability. These results are consistent with the previous theoretical articulations (e.g., Gulliksen, 1945; Symonds, 1928) and empirical findings (e.g., Jang, 2009; X. Wang et al., 2019) that larger variability and extreme levels of item difficulty reduce test score variances and thus lead to lower reliability and classification accuracy of test scores. As reliability is essentially a correlation of test with itself and correlations decrease with reduced score variability, larger variability and extreme levels of item difficulty are likely to result in decreased subscore reliability.

Importantly, the high mean condition in this study also featured the mismatch between average item difficulty and cut-scores. Test information is typically maximized at the center of item difficulty distribution and hence, it is often desired to set cut-scores at the center of item difficulty distribution with a smaller variability. Therefore, setting cut-scores where the conditional standard error of measurement is large, as in high mean conditions, is likely to result in lower classification accuracy (Lathrop & Cheng, 2013). Note that although both extreme

means and larger variability of item difficulty negatively affected classification accuracy, the extreme item difficulty exerted greater detrimental effect. This is because alignment between average item difficulty and average examinee ability affects base mastery rates which in turn have a huge impact on classification accuracy.

Research Question 2: Subscore Distinctiveness of Diagnostic Methods

In this section, findings related to subscore distinctiveness (research question 2) are discussed. This section is organized into two parts, comparative subscore distinctiveness of diagnostic methods and the impact of various factors (i.e., L, D, and R) on subscore distinctiveness.

Comparative Subscore Distinctiveness of Diagnostic Methods

In this study, subscore distinctiveness was examined based on intercorrelations of subscores as well correlations between subscores and overall scores. These correlations were computed using both discrete and continuous subscores: mastery categories and proficiency scores. The most notable finding regarding the subscore distinctiveness was that inter-subscore correlations and correlations between subscores and overall score estimates were lower for ADCM compared to those of IRT-based methods (M = .70, 69, and .60 for inter-subscore correlations of AIRT, MIRT, and ACDM, respectively; M = .91, 91, and .81 for correlations between overall scores and subscores of AIRT, MIRT, and ACDM, respectively) when they were computed based on continuous proficiency scores. As noted earlier, those lower correlations of ACDM are in part attributable to a rather restricted range of ACDM proficiency scores that are probability estimates ranging from 0 to 1. However, correlations based on logit transformed ACDM proficiency scores (M = .64 for inter-subscore correlation; M = .88 for correlations between overall score estimates and subscores) were still lower than those of IRT-based methods.

Therefore, these results call for reviews on how DCMs versus IRT-based methods estimate subscores.

AIRT and MIRT enhance subscores by borrowing information from out-of-scale items (i.e., the examinee's performance on other subdomains in the test), thus, resulting in subscores that are more highly correlated when compared to raw subscores or subscores estimated from the unidimensional approach. As Bulut et al. (2016) indicated, AIRT and MIRT "may alter the factor composition of what is measured by subscores, increasing variance attributable to the general factor common to all subtests and decreasing the contribution of the subtest-specific factors" (p. 13). The relatively lower correlation of DCM estimates also stems from differences in theoretical assumptions between DCMs and IRT-based methods. DCMs assume a discrete rather than continuous latent variable and they are designed to provide examinees with more fine-tuned profiles of discrete attributes rather than profiles of point estimates measured on a continuous continuum.

Due to these differences, subscores estimated in continuous scales are likely to be more highly correlated for IRT-based methods than for DCMs although all three methods share the commonality that they incorporate correlations among subscores into the estimation procedure. Conversely, when subscore distinctiveness is evaluated using binary mastery categories, differences in point estimates of subscores largely diminish, hence leading to comparable results across diagnostic methods as shown in this study (M = .70, .68, and .70 for inter-subscore correlations of AIRT, MIRT, and ACDM, respectively; M = .91, .90, and .91 for correlations between overall score estimates and subscores of AIRT, MIRT, and ACDM, respectively).

Impact of Various Factors on Subscore Distinctiveness

Both inter-subscore correlations and correlations between overall scores and subscores tended to increase as subscales got shorter although such a pattern was more pronounced for

inter-subscore correlations. Expectedly, both inter-subscore correlations and correlations between subscores and overall score estimates also increased as true inter-subscale correlations increased. Besides, intercorrelations of subscores tended to slightly increase as item difficulty distribution deviates from examinee ability distribution, whereas correlations between overall scores and subscores tended to be comparable across item difficulty distributions. All these patterns were consistent regardless of the type of subscores.

Note that shorter subscales and item difficulty distribution deviated from examinee ability distribution involve relatively large measurement errors and small test score variances, hence, yielding subscores that are less precise and stable. As noted by Choi and Papageorgiou (2020), the distinctive contribution of each subscore can constitute meaningful signals rather than noise only when subscores are accurate and stable enough. Thus, those conditions are unlikely to yield distinct subscores. Furthermore, given that lower reliability leads to a reduction in the size of a correlation, under those conditions, intercorrelations of subscores would be even higher after being corrected for attenuation. As noted earlier, researchers (Haladyna & Kramer, 2004; McPeek et al., 1976) suggested that if disattenuated inter-subscore correlations are larger than .90, the subscores are not considered distinct from each other and therefore not worth reporting.

Research Question 3: Incremental Criterion-Related Validity Evidence of Diagnostic Methods

In this section, findings related to incremental criterion-related validity evidence (research question 3) of subscores are discussed. This section is organized into two parts: comparative incremental criterion-related validity evidence of diagnostic methods and the impact of various factors (i.e., L, D, and R) on incremental criterion-related validity evidence of subscores.

Comparative Incremental Criterion-Related Validity Evidence of Diagnostic Methods

Across all conditions, all diagnostic methods explained a comparable amount of criterion variation over and above overall scores. Although IRT-based methods explained slightly more variation than did ACDM, their differences were statistically insignificant. Specifically, IRT-based methods and ACDM on average added about 0.8% and 0.6% to the variance accounted for by overall scores alone when designated criterion validity was 64% shared variance. In contrast, IRT-based methods and ACDM on average added about 0.3% and 0.2% to the variance accounted for by overall scores alone when designated criterion validity was 25% shared variance. Thus, although subscore contribution over and above overall scores was statistically significant, the amount subscores add to the criterion variance explained by overall scores alone may be practically trivial in both high and low criterion validity coefficient conditions.

These results are consistent with empirical findings by Kunina-Habenicht et al. (2017) who investigated validity evidence of a newly developed, diagnostic arithmetic assessment (DAA) for elementary school students. Authors examined the extent to which proficiency scores of DCM explain additional variance in school grades in mathematics and unidimensional proficiency scores for a national standards-based assessment for mathematics over and above proficiency scores from UIRT. In this study, DCM scores were moderately correlated with both criterion variables: correlations ranged from -.41 to -.53 for the school grades in math and ranged from .41 to .48 for the unidimensional scores for the national math assessment. In contrast, scores from UIRT were more highly correlated with those two criterion variables (i.e., correlations were -.65 and .61). Besides, diagnostic subscores were relatively distinct from each other (i.e., intercorrelations of subscores ranged from .23 to .68) and from proficiency scores of UIRT (i.e., correlations between diagnostic subscores and unidimensional scores ranged from.61, to .81). Nevertheless, Kunina-Habenicht et al. (2017) found only a negligible amount of incremental validity evidence of four attribute scores (i.e., proficiency scores of DCM explained only 0.3%

and 1.6% of additional variance of school grades in math and of the national math assessment over and above proficiency scores from UIRT, respectively).

Note that in the present study, two subscales were generated with equal length based on the same parameter distribution specifications. Also, criterion variables were created to have much the same relationship with two subscores given that the primary interest of the study was whether subscores add to the variance accounted for over overall scores rather than whether the pattern of subscores is related to the criterion. Consequently, two subscores contributed similarly to the prediction of criterion variables across conditions although the relative contribution of two subscores versus overall scores changed as a function of manipulated factors.

In operational settings, the extent to which each subscore differentially contributes to the prediction of criterion variables may vary greatly depending on the type of criterion variables and multidimensional tests employed. For example, Kunina-Habenicht et al. (2017) used the criterion variable (i.e., unidimensional proficiency scores from national assessment of math) representing broader construct covering all core competencies of mathematical literacy, whereas diagnostic subscores represented four refined elementary skills in arithmetic (i.e., "addition/subtraction", "multiplication/division", "modeling skills", and "skills for using measurement units"). Therefore, as expected, each subscore was similarly correlated to the criterion variable. However, in situations where dimensions of diagnostic tests may be broadly defined as composite abilities, such as math or verbal ability (e.g., Quantitative and Verbal scores of the SAT), contribution of each subscore for the prediction of unidimensional criterion variables may be more likely to differentiated from each other, yielding a significant criterion-related pattern. The current study design more closely resembles the former situation given that DCMs primarily aim to provide more fine-tuned diagnostic feedback for test users.

Impact of Various Factors on Incremental Criterion-Related Validity Evidence of Subscores

Overall predictability increased as subscales get longer, item difficulty distribution gets more aligned with examinee ability and cut-scores, and intercorrelations of subscores increased. These results are anticipated given that these conditions are associated with higher subscore reliability/classification accuracy. As reliability is in essence correlation of a test with itself, if a test correlates highly with itself, it is more likely to correlate highly with another variable (Goodwin & Leech, 2006). Therefore, under those conditions, subscores were more highly correlated with criterion variables, hence yielding higher predictive validity evidence. In addition, other things being equal, overall predictability increased as the squared correlation between subscores and criterion variables increased given a direct impact of the criterion validity evidence on the predictability of subscores.

However, the unique contribution of subscores over and above overall scores (i.e., $\Delta R^2_{M_2-M_1}$) greatly decreased as intercorrelations of subscores increased and criterion validity coefficients decreased. Average $\Delta R^2_{M_2-M_1}$ of MIRT was 2.3%, 0.6%, 0.2, and 0.1% when inter-subscore correlations were .30, .50, .70, and .80 under high criterion validity coefficient condition. In contrast, the respective $\Delta R^2_{M_2-M_1}$ was 0.9%, 0.3%, 0.1%, and 0% under low criterion validity coefficient condition. Therefore, the upper bound of inter-subscore correlations for subscores to have incremental criterion-related validity evidence over and above the overall score estimate seemed to be .80 and .70 for high and low criterion validity coefficient conditions, respectively based on the current study design. This result is consistent with the previous findings that subscore added value is primarily a function of subscore distinctiveness.

Summary of Findings

In general, DCM yielded classification accuracy lower than that of IRT-based methods. However, differences between DCM versus IRT-based methods were not practically meaningful when average item difficulty was matched with average examinee ability and cut-scores. In

contrast, when average item difficulty was mismatched with average examinee ability and cut-scores, DCM yielded classification accuracy significantly different from that of IRT-based methods, especially when classification accuracy was measured by sensitivity and specificity. Compared to IRT-based methods, DCM also yielded subscores that are more distinct from each other and from overall scores when subscores were estimated on continuous scales. Despite the greater subscore distinctiveness, the unique contribution of DCM estimates to the prediction of criterion variables over and above overall scores was slightly smaller than that of IRT-based methods given that DCM estimates did not correlate with criterion variables as highly as did subscores from IRT-based methods. Although such differences were not practically meaningful, the contribution of subscores to the prediction of criterion variables over and above overall scores was statistically significant only for IRT-based methods when squared correlations between subscores and criterion variables were low (i.e., 25% shared variance) due to the relatively small effect sizes. In contrast, when squared correlations between subscores and criterion variables were high (i.e., 64% shared variance), the unique contribution of subscores was statistically significant for both DCM and IRT-based methods.

In addition, classification accuracy was higher for longer subscales, item difficulty distribution more aligned with examinee ability distribution and cut-scores, and higher intercorrelations of subscores. On the other hand, subscore distinctiveness was greater for longer subscales, item difficulty distribution more aligned with examinee ability distribution and cut-scores, and lower intercorrelations of subscores. In contrast, incremental criterion-related validity evidence of subscores primarily increased as intercorrelations of subscores decreased and criterion validity coefficients increased.

Limitations of the Present Study and Future Research Directions

As is the case for any simulation study, the results of this study may only be generalizable to testing conditions manipulated in this study. For example, in DCMs, the accuracy of estimation and classification is generally a function of the number of attributes, data structure, sample sizes, and complexity of analysis models (Ravand & Baghaei, 2019). Therefore, studies based on different combinations of these factors may yield results different from those of the present study. Besides, irregularities involved in operational data may lead to different results even under similar testing conditions. Therefore, future research needs to evaluate if the results of this study are replicated using both simulated and operational data that involve different values of study design elements.

To mimic more common testing situations where tests are constructed based on a continuous measurement model, this study generated data based on the two-dimensional IRT models. However, to be of optimal use, DCMs must be applied in conjunction with CDA which is designed for providing diagnostic feedback at a finer conceptual cognitive grain size. Therefore, the discrepancy between data generating and data analysis models might put ACDM in a relatively disadvantageous position. In the future study, it would be ideal to perform simulations using both DCM and IRT as data-generating models and compare two results.

Also, the present study evaluated the added value of subscores using continuous proficiency scores only. When the focus of the study is to evaluate the added value of subscores concerning classification, one can consider the procedure proposed by Sinharay (2014). Similar to Haberman's (2008) added value analysis, Sinharay's procedure determines subscore added value if examinee classifications based on observed subscores better predict examinees' classifications based on true subscores than does examinee classifications based on observed total scores. This procedure involves grouping examinees based on one or more cut-scores.

As in the majority of prior studies, this study generated data assuming simple rather than complex structure given that adding complex structure items to the subtests leads to a decrease in the diagnostic value and orthogonality of subscores (Feinberg & Wainer, 2014; Madison & Bradshaw, 2015). Nevertheless, the complex items may still be meaningful as parts of the overall test score in some situations. Especially, in the CDA approach, the assessment is often designed to measure complex constructs that may not be easily separated and measured in isolation. Such assessment has become increasingly more prevalent over the past decade targeting skills related to complex problem solving, digital proficiencies, creativity, and computer and information literacy (Ercikan & Oliveri, 2016; Svetina et al., 2017). Therefore, more research comparing performances of diagnostic methods under the complex structure is warranted.

Although it was beyond the scope of this study, it would be interesting to investigate the invariance of incremental criterion-related validity evidence between subgroups of examinees. Researchers (e.g., Rios & Miranda, 2021; Sinharay & Haberman, 2014) have demonstrated that subscores had added value for some test takers but not for others. Therefore, for future research, it is suggested to examine the extent to which added value of subscores evaluated in terms of incremental criterion-related validity evidence varies depending on various subsample characteristics, such as average ability, variability of ability, and intercorrelations of subdomain abilities.

Conclusion and Recommendations

Based on findings from this study, AIRT or MIRT would be in general preferable over DCM as diagnostic means when item responses are obtained from IRT-based assessment forms. Additionally, for classification purposes it would be advantageous to have more items in each subscale, match average item difficulty with examinee ability and cut-scores, concentrate items around the cut-score with a smaller variability, and use subscores that are highly intercorrelated.

If average item difficulty of tests in use is mismatched with average examinee ability and cut-scores, test users must ensure adequate classification accuracy for individual attributes or attribute vectors using additional information such as out-of-test collateral information (e.g., demographics and other educational achievement variables) that can supplement test data (Deonovic et al., 2019). On the other hand, to ensure that subscores add to the criterion variance accounted for by overall scores alone, it would be critical to obtain subscores from assessments designed for diagnostic purposes so that subscores are not too highly correlated to each other and to overall scores. Other things being equal, the upper bound of inter-subscore correlations to support the incremental criterion-related validity evidence of subscores would depend on the magnitude of the criterion validity coefficient; the higher the criterion validity coefficient becomes, the higher intercorrelations subscores can afford.

Bibliography

Ackerman, T. A., Gierl, mark J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51. http://www.ncme.org.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

Almond, R. G., Dibello, L. V, Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*(4), 341–359.

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

Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp. 159–172). Boca Raton, FL: Chapman & Hall.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and Psychological Measurement*, *79*(1), 65–84. https://doi.org/10.1177/0013164418763255

Bradshaw, L., Izsák, A., Templin, J. L., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. https://doi.org/10.1111/EMIP.12020

Brennan, R. L. (2012). *Utility indexes for decisions about subscores*. CASMA Research Report No. 33. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models*. [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations Publishing.

Bulut, O., Davison, M. L., & Rodriguez, M. (2016). Estimating between-person and within-person subscore reliability with profile analysis. *Multivariate Behavioral Research*, *0*(0), 1–19. https://doi.org/10.1080/00273171.2016.1253452

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chan, C.-K. (2006). *An exploratory study of predictive validity and diagnostic utility of cognitive profile patterns on the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R) clinical database*. [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations Publishing.

Chen, P. H. (2006). The influences of the ability estimation methods on the measurement accuracy in multidimensional computerized adaptive testing. *Bulletin of Educational Psychology*, *38*(2), 195–211.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika*, *82*(3), 660–692. https://doi.org/10.1007/s11336-016-9545-6

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Chin, T.-Y. (2011). *Accuracy and robustness of diagnostic methods: Comparing performance across domain score, multidimensional item response, and diagnostic categorization models*. [Docotral dissertation, University of Nebraska-Lincoln]. ProQuest Dissertations Publishing.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633–665. https://doi.org/10.1007/S11336-009-9125-0

Chiu, C.-Y., & Seo, M. (2009). Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLs reading assessment. In *IERI monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 137–159). IEA-ETS. www.ierinstitute.org

Choi, I., & Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: A case of reading and listening subscores for young EFL learners. *Language Testing*, *37*(2), 254–279. https://doi.org/10.1177/0265532219879654

Clay, M. (1998). *By different paths to common outcomes*. York, ME: Stenhouse.

Clay, M. (2002). *An observational survey of early literacy achievement*. Auckland, New Zealand: Heinemann.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York, NY: Plenum Press.

Close, C. N. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: combining theory with data* [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations Publishing.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cui, Y., Gierl, M., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19–38.

Culpepper, S. A. (2009). A multilevel, nonlinear profile analysis model for dichotomous data. *Multivariate Behavioral Research*, *44*, 646–667.

da Silva, M. A., Liu, R., Huggins-Manley, A. C., & Bazán, J. L. (2019). Incorporating the Q-matrix into multidimensional item response theory models. *Educational and Psychological Measurement*, *79*(4), 665–687. https://doi.org/10.1177/0013164418814898

Davison, M. L., Chang, Y.-F., & Davenport, E. C. (2014). Modeling configural patterns in latent variable profiles: association with an endogenous variable. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 81–93. https://doi.org/10.1080/10705511.2014.859507

Davison, M. L., & Davenport, E. C. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods*, *7*(4), 468–484.

Davison, M. L., Davenport, E. C., Chang, Y. F., Vue, K., & Su, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement*, *52*(3), 263–279. https://doi.org/10.1111/jedm.12081

Davison, M. L., Kim, S.-K., & Close, C. (2009). Factor analytic modeling of within person variation in score profiles. *Multivariate Behavioral Research*, *44*(5), 668–687. https://doi.org/10.1080/00273170903187665

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and application. *Journal of Educational Measuremen*, *45*(4), 343–362.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9214-8

de la Torre, J., Andries van der Ark, L., & Rossi, G. (2018). Analysis of clinical data from a

cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296.
de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8
de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640
de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction substraction data. *Psychometrika*, *73*(4), 595–624.
de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227–249. https://doi.org/10.1111/j.1745-3984.2010.00110.x
de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*. *47*(1), 115-127.
de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355–373.
de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89–97. https://doi.org/10.1016/j.pse.2014.11.001
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), 62.
de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*(4), 296–316. https://doi.org/10.1177/0146621610378653
DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069
Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 437–460). Springer, Cham.
Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.
Desmarais, M., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedigns of the 6th International Conference on Artificial Intelligence in Education (pp. 441–450).* Heidelberg: Springer.
DiBello, L., Stout, W., & Roussos, L. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979–1030). Amsterdam, Netherlands: Elsevier.
Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). *A comparison of subscale score augmentation methods using empirical data*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
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 Fall*,

31(3), 241–259.
Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer-Verlag.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *PSYCHOMETRIKA*, *78*(1), 14–36. https://doi.org/10.1007/S11336-012-9296-Y
Ercikan, K., & Oliveri, M. O. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education*, *29*, 310–318.
Feinberg, R. A. (2012). *A simulation study of the situations in which reporting subscores can add value to licensure examinations*. [Doctoral dissertation, University of Delaware]. ProQuest Dissertations Publishing.
Feinberg, R. A., & Wainer, H. (2014). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational Measurement: Issues and Practice*, *33*(3), 47–54.
Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*, 10–26.
Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*(1), 67–82. https://doi.org/10.1177/0146621610367787
Firestone, W. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*, *43*(2), 100–107. https://doi.org/10.3102/0013189X14521864
Fu, J., & Qu, Y. (2018). *A review of subscore estimation methods*. ETS Research Report (RR-18-17). Princeton, NJ: ETS. https://doi.org/10.1002/ets2.12203
Garcia, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *26*, 372–377.
Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 293–313.
Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretative guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220.
Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(1), 732–764.
Goodman, Y., Watson, D., & Burke, C. (2005). *Reading miscue inventory: From evaluation to instruction*. Katonah, NY: Owen.
Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, *74*(3), 249–266.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.
Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, *10*(2), 79–91.
Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636
Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item

response theory. *Psychometrika*, *75*(2), 209–227. https://doi.org/10.1007/S11336-010-9158-4

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *The British Psychological Society*, *62*, 79–95. https://doi.org/10.1348/000711007X248875

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 1031–1038). Amsterdam: Elsevier North-Holland.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions*, *27*(4), 349–368.

Harris, D. J., & Hanson, B. A. (1991). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. [Doctoral dissertation, University of Illinois Urbana-Champaign], Proquest Dissertation Publishing.

Hartz, S. M., Roussos, L., & Stout, W. (2002). *Skills diagnosis theory & practice*. User Manual for Arpeggio software [Computer software manual]. Princeton, NJ: Educational Testing Service.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2). http://www.copyright.com/

Henson, R. A., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277. https://doi.org/10.1177/0146621604272623

Henson, R. A., & Templin, J. L. (2008). *Implementation of standards setting for a geometry end-of-course exam*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

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

Huang, L. (2015). *Improving the use of subscores on a test battery: Some reliability and va- lidity evidence from the Wechsler Intelligence for Children Scale.* [Doctoral dissertation, University of Minnesota], Proquest Dissertation Publishing.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, *15*(1).

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 31–73. https://doi.org/10.1177/0265532208097336

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*, 359–383.

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y.-H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning*, *63*, 400–436.

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement Winter*, *55*(4), 635–664. https://doi.org/10.1111/(ISSN)1745-3984

Johnson, M. S., & Sinharay, S. (2020). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*, *45*(1), 5–31. https://doi.org/10.3102/1076998619864550

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, *14*(1), 49–72. https://doi.org/10.1080/15305058.2013.835728
Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, *28*(6), 407–426.
Kelly, T. L. (1927). *The interpretation of educational measurement*. New York: World Book.
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing*, *17*(4), 277–301. https://doi.org/10.1080/15305058.2017.1291517
Kunina, O., Rupp, A. A., & Wilhelm, O. (2008). *Convergence of skill profiles for cognitive diagnosis models and other multidimensional scaling approaches: An empirical illustration with a diagnostic mathematics assessment*. Presented at the annual meeting of the Psychometric Society (IMPS), Durham, NH.
Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. Log & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 610–627). Walden, MA: Wiley-Blackwell.
Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, *37*(3), 226–241. https://doi.org/10.1177/0146621612471888
Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. https://doi.org/10.1080/15434300902985108
Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, *40*(6), 405–417.
Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 17–46.
Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–564. https://doi.org/10.1177/0146621612456591
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599
Longabach, T., & Peyton, V. (2018). A comparison of reliability and precision of subscore reporting methods for a state English language proficiency assessment. *Language Testing*, *35*(2), 297–317. https://doi.org/10.1177/0265532217689949
Luecht, R. M. (2003). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.
Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (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.-E. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, and Evaluation*, *14*(4), 1–10.
Ma, W., & de la Torre, J. (2020). Gdina: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14
Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Article Applied Psychological Measurement*, *40*(3), 200–217. https://doi.org/10.1177/0146621615621717
Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511. https://doi.org/10.1177/0013164414539162
Maris, E. (1995). Psychometric latent response model. *Psychometrika*, *60*, 523–547.
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.
Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*(2), 209–228. https://doi.org/10.1007/BF02294836
Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*, 71–101.
McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE advanced psychology test*. GRE Board Professional Report No. 74–4P. Princeton, NJ: Educational Testing Service.
Min, S., & He, L. (2021). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing*, 1–27. https://doi.org/10.1177/0265532221995475
Montero, D. H., Monfils, L., Wang, J., Yen, W. M., & Julian, M. W. (2003). *Investigation of the application of cognitive diagnostic testing to an end-of-course high school examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. https://www.researchgate.net/publication/237276515
National Joint Committee on Learning Disabilities. (2011). Comprehensive assessment and evaluation of students with learning disabilities. *Learning Disability Quarterly*, *34*, 3–16.
Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
Nye, C. D., Joo, S.-H., Zhang, B., & Stark, S. (2020). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486. https://doi.org/10.1177/1094428119833158
Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447.
Papageorgiou, S., & Choi, I. (2018). Adding value to second-language listening and reading subscores: Using a score augmentation approach. *International Journal of Testing*, *18*(3), 207–230.
Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure Exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266–285. https://doi.org/10.1080/08957347.2010.486287
Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 1–18. https://doi.org/10.1177/0734282915623053
Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments,

practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278
Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.
Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36.
Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. http://www.springer.com/3463
Rios, J. A., & Miranda, A. A. (2021). What are the conditions associated with subscore added value noninvariance? Implications for improving subscore interpretation fairness. *Educational Measurement: Issues and Practice*, *40*(1), 69–78.
Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*(3), 25–38.
Roussos, L. A., Templin, J. L., & Henson, R. (2005). *Theoretically grounded linking and equating for mastery/non-mastery skills diagnosis models.* Unpublished ETS Project Report, Princeton, NJ: Educational Testing Service.
Rupp, A. A., & Templin, J. L. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545
Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-Art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866
Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. New York, NY: Guilford.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology*, *11*, 621251. https://doi.org/10.3389/fpsyg.2020.621251
Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104
Sijtsma, K., & Junker, B. W. (2006). Item response theory: past performance, present developments, and future expectations. *Behaviormetrika*, *33*(1), 75–102.
Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement Summer*, *47*(2), 150–174.
Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement Summer*, *51*(2), 212–222. https://doi.org/10.1111/(ISSN)1745-3984
Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey*. ETS Research Memorandum No. 08–18. Princeton, NJ: Educational Testing Service.
Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement*, *7*(1), 46–49.
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. https://doi.org/10.1080/15305058.2013.822712
Sinharay, S., Haberman, S. J., & 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., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of*

Diagnostic Classification Models (pp. 359–377). Cham, Switzerland: Springer International Publishing.
Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573. https://doi.org/10.1080/00273171.2010.483382
Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice Fall*, *30*(3), 29–40.
Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2019). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score Reporting Research and Applications* (pp. 35–49). New York, NY: Routledge.
Skaggs, G., Wilkins & Serge, J. L. M., & Hein, F. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing*, *16*(4), 310–330. https://doi.org/10.1080/15305058.2016.1145683
Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 63–86.
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrik*, *52*, 589–617.
Svetina, D., Valdivia, A., Underhill, S., Dai, S., & Wang, X. (2017). Parameter recovery in multidimensional item response theory models under complexity and nonnormality. *Applied Psychological Measurement*, *41*(7), 530–544. https://doi.org/10.1177/0146621617707507
Symonds, P. M. (1928). Factors influencing test reliability. *Journal of Educational Psychology*, *19*(2), 73.
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, MN: University of Minnesota.
Tate, R. L. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159–203.
Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*(2), 89–112.
Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275. https://doi.org/10.1007/s00357-013
Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287
Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*(7), 559–574. https://doi.org/10.1177/0146621607300286
Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 29–40). New York, NY: Springer.
Thissen, D., & Wainer, H. (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates

Publishers.
Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervalsfor effect sizes. *Psychology in the Schools*, *44*(5), 423–432.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.
Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339–368.
Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (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). Hillsdale, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9781410604729-16
Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.
Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*(2), 119–134. https://doi.org/10.1177/0146621614545983
Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116
Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and aattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement Winter*, *52*(4), 457–476. https://doi.org/10.1111/(ISSN)1745-3984
Wang, X., Svetina, D., & Dai, S. (2019). Exploration of factors affecting the added value of test subscores. *The Journal of Experimental Education*, *87*(2), 179–192.
Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479–494.
Wilhelm, O., & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 53–57.
Wyse, A. E. (2011). The potential impact of not being able to create parallel tests on expected classification accuracy. *Applied Psychological Measurement*, *35*(2), 110–126.
Xin, T., & Zhang, J. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement*, *39*(1), 44–61.
Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*(3), 625–649.
Xu, X., & von Davier, M. (2008). *Linking for the general diagnostic model*. ETS Research Report. Princeton, NJ: ETS.
Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*(3), 339–360.
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. https://doi.org/10.1177/0146621606291559
Yen, W. M. (1987). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.
Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.
Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its

application to approximate simple structure. *Psychometrika*, *64*, 213–249.