LUO, XIAO, Ph.D. The Optimal Design of the Dual-purpose Test. (2013)
Directed by Dr. Richard M. Luecht. 155 pp.

Traditional test development focused on one purpose of the test, either ranking test-takers or providing diagnostic profiles for test-takers. Embedding both the ranking and diagnostic purposes in one assessment instrument would be a great advancement to the test functionality and utility. Our understandings regarding how such dual-purpose test should be optimally design and analyzed, however, were dwarfed by the growing needs for it in practice. Potential psychometric challenges related to the dual-purpose testing were not fully addressed in the literature. The present study provided a systematic comparison of various plausible designing and analyzing paradigms for the dual-purpose test in conditions with varying test length and dimensionality of true abilities.

Results suggested that in order to obtain accurate and reliable total score and subscores, the test should be designed with multidimensionality and at least 10 items per domain and analyzed using the multidimensional IRT model. Specifically, the unidimensional dual-purpose test was able to produce reliable and accuracy but not diagnostically meaningful scores. Subscores obtained from an essentially unidimensional test were either unable to provide added value to the total score according to the PRMSE criterion or homogeneous to each other according to disattenuated correlations. The idiosyncratic multidimensional design was able to yield accurate, reliable, and diagnostically useful scores, but the validity of the diagnostic subscores was questionable, whose correlation disagreed with the true correlational structure. Consequently, even though subscores were identified distinct from the total score according to the PRMSE

criterion, they were still nearly identical to each other according to the disattenuated correlations. On the other hand, the principled multidimensional design showed slightly lower accuracy and reliability in scores due to the principled “simple structure” of test design, but this sacrifice of accuracy and reliability ensured the interpretability and validity of diagnostic subscores, whose empirical correlational structure approximated the true structure.

Furthermore, with respect to calibration methods, unidimensional calibration was found failing to distinguish subscores, and thus failing to give subscores useful diagnostic information, even though the subscores sometimes appeared more accurate and reliable than those obtained with the other two calibrations. The confirmatory multidimensional calibration and separate unidimensional calibration delivered very comparable results. Finally, alternative scoring methods were found either inappropriate to use or offering insignificant improvements over the raw scores.

THE OPTIMAL DESIGN OF THE DUAL-PURPOSE TEST

by

Xiao Luo

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2013

Approved by

____
Committee Chair

To Xinrui and Sophie.

APPROVAL PAGE

This dissertation written by XIAO LUO has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ____

Committee Members ____

____

____

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

It has been a long tradition to design tests that serve one purpose. Mostly, the primary purpose of an assessment instrument is to quantify the ability being measured (i.e., construct) and thereby rank test-takers according to their numeric scores on the latent scale. Scores that test-takers receive summarize their performances in the testing domain and possibly predict their future performance in a more generalized domain. For this ranking purpose, test scores can be used a as an important evaluative criterion to assist in critical decision-making processes, such as granting admission, awarding scholarship, assigning class placement, certificating or licensing profession, and so forth. In educational settings, for instance, a student's SAT® or GRE® score is commonly referred to as an indicator of the student's mastery of skills and knowledge (i.e., score of an achievement test) or a predictor of the student's academic success in higher education (i.e., score of an aptitude test) in the process of granting college or graduate school admission and/or awarding scholarships. Although numeric scores are not always directly reported to test-takers in professional certification and licensure settings, each test-take is still scored with a total score, indicating the test-taker's preparedness of entering into the profession. The pass/fail decision will be made by comparing the total score to a predetermined passing standard set to mirror the skills and knowledge required by the safe and effective entry-level practice in that profession. In addition to the ranking purpose, some other tests are designed for the diagnostic purpose. The fine-grained diagnostic scores offer important information that is missing in the total score. For instance, two test-takers who receive the same total scores may show different patterns of diagnostic scores, in other words, distinct patterns of strengths and weaknesses. Diagnostic subscores thereby become useful in further distinguishing these two test-takers, where the total score regard them as indistinguishable. The No Child Left Behind Act of 2001 even mandated educational tests to report diagnostic scores in order to facilitate educators addressing the student's specific academic needs (Sinharay, 2010). Although diagnostic report is not mandatory for professional certification and licensure tests, some testing programs still elect to provide, especially to failing test-takers, diagnostic subscores in hopes of guiding test-takers' preparations for the examination and profession in the future.

It would be a desirable extension for the test, which was originally designed to report the total score alone, to report diagnostic subscores as well. The total score is therefore reported for the test's primary purpose, or function, of quantifying and ranking test-takers, whereas subscores are reported for the secondary purpose, which is to diagnose test-takers' strengths and weaknesses. Accordingly, the test of this kind is defined as the dual-purpose test in the present study. It may be deemed, outside of psychometric communities, as natural for a test to serve more than one purposes; however, inside of psychometric communities it is well acknowledged that many psychometric difficulties would actually arise if a test has, by design, more than one purposes of score use. Per Wainer et al. (2001), the ranking and diagnostic purposes were “antithetic” in the sense that one required the test to be focused on a narrowly-defined and coherent domain whereas the other required the test to encompass broad and distinctive domains. That is, in formal terminology, the ranking purpose requires the complete test to be unidimensional, with all items in all content domains consistently measuring the same construct, so that the total score of the test serves as a reliable and valid measurement of the overall ability. The diagnostic purpose, on the other hand, requires the test to be multidimensional on the premise that the overall ability further comprises of multiple domain-specific abilities, which are measured in individual subtests or dimensions. Without multidimensionality, the test may fail to produce meaningful and useful diagnostic subscores, since items are too akin to yield distinct information from the total score (Haberman, 2008; Luecht, Gierl, Tan, & Huff, 2006; Wainer et al., 2001). In short, a successful dual-purpose test purports to be unidimensional and multidimensional simultaneously.

Another psychometric difficulty is related to the test reliability. To validate a dual-purpose test, both the total score and subscores need to be supported by theoretical and empirical evidences of high reliability (American Educational Research, American Psychological, & National Council on Measurement in, 1999). Fundamentally, the reliability is defined as the consistency of a test-taker’s scores over multiple administrations. The simplest incarnation of this definition is the test-retest reliability, which is given by the correlation between two administrations of the same test spanned over a period of time that is long enough to erase memorization of the test content and short enough to avoid growth of the ability. Because it is typically impossible to administrate the same test repetitively in practice, even just twice, estimators that approximate reliability using one administration were developed to circumvent this problem. Among all estimators, Cronbach’s alpha, also known as the KR-20 or internal consistency, is arguably the most common and useful one in practice, which estimates the degree to which items within the test are measuring the same construct statistically by the ratio of the between-item covariance to the total score variance (Cronbach, 1951; Kuder & Richardson, 1937). According to the Spearman-Brown prophecy formula, test reliability grows in proportion to the increase of test length. That means, while the complete test has an acceptably high reliability, subtests may not achieve good reliabilities due to the insignificantly shorter test lengths. The reliability issue is further complicated by the dimensionality. If the test were designed with multidimensionality to yield meaningful diagnostic subscores from subtests, the “internal consistency” of the test would be contaminated by such multidimensionality. Therefore, the complete test is unlike to show high reliability, even though subscore may have been satisfactorily reliable.

The last psychometric difficulty is concerning which analysis paradigm should be used to calibrate the dual-purpose test and score test-takers. Unidimensional item response theory (UIRT) was developed with the purpose of producing psychometrically sound total score, but not subscores. Alternative subscoring methods were proposed to be used in addition to the UIRT analyses to derive subscores (de la Torre & Patz, 2005; de la Torre & Song, 2009; de la Torre, Song, & Hong, 2011; Wainer et al., 2001; Yen, 1987). In contrast, multidimensional item response theory (MIRT) was developed to report domain-specific subscores, leaving the total score of the overall ability be computed with alternative methods (Graybill & Deal, 1959; Longford, 1997; Luecht et al., 2006). There is little research in literature specifically comparing calibration and scoring methods in terms of their effectiveness of deriving both good total scores and subscores. Much is still unknown about the optimal analysis paradigm for the dual-purpose test.

Given considerations above, it is obviously seen that embedding two “antithetic” purposes of score use into one test would bring forth several psychometric difficulties for the design and analysis of the test. Any naïve treatment of these issues could lead to a flawed test with misleading scores and unfulfilled purposes. For example, empirical studies found that subscores of many existing dual-purpose tests failed to provide added value to the total score, and thus were not worth reporting (Haberman, 2008; Puhan, Sinharay, Haberman, & Larkin, 2010; Sinharay, 2010; Sinharay, Haberman, & Puhan, 2007). Therefore, the purpose of the present study was to probe for the optimal design of the dual-purpose test and the effective scoring procedure that should be coupled with that design. The performances of various plausible designs and analysis paradigms were compared across conditions with varying test length and dimensionality of true abilities. Findings of this study were expected to advise practitioners on how to design and analyze a dual-purpose that contains both valuable total score and diagnostically meaningful subscores.

CHAPTER II

REVIEW OF THE LITERATURE

Unidimensionality

It has been a long tradition for practitioners to design, develop, and administer unidimensional tests, in which all items in all content domains measure the same construct. A test-taker's proficiency level of the construct, in return, is considered as the single determinant of his or her test performance. The construct can be a simple latent ability (e.g., the ability of summation operation) or a complex combination of multiple latent abilities (e.g., the ability of arithmetic operation, consisting of abilities pertaining to summation, subtraction, multiplication, and division), as long as the composition of these specific abilities is consistent throughout the test. The total score obtainable from the test is thus reported as a summary of the test-taker's overall test performance and the latent ability. The total score also serves as a piece of very interpretable and usable information for ranking test-takers and making decisions.

To illustrate the concept of unidimensionality, consider the true score model in the classical test theory (CTT). As the model conceptualizes, the observed score consists of two unobservable components, the true score and the error:

$$x = \tau + \varepsilon \tag{2.1}$$

where x is the observed score, τ is the true score, and ε is the error term (Allen & Yen, 1979). The true score is the test-taker's true ability of the construct being measured. The error term is an unwanted but inevitable variation added to the true score, which can be caused by a variety of contextual variables. In general, the error term is assumed to have a standard normal distribution *N(0, 1)*, meaning it would be summed to zero and cancelled out of the equation in the long term. Regardless of whether the construct is simple or composite one, the test-taker's ability is packaged into and represented by one parameter, namely the true score τ. This can be considered as the incarnation of the unidimensionality assumption in the true score model.

The same is true for UIRT models. Take the famous unidimensional 3-parmaeter logistic (3PL) model for example. The probability of getting a correct response to an item in the 3PL model is given by

$$P\left(u_{ij} = 1 \mid \theta_i, a_j, b_j, c_j\right) = c_j + \frac{1-c_j}{1+\exp\left[-1.7a_j\left(\theta_i - b_j\right)\right]} \tag{2.2}$$

where u_{ij} is test-taker i's response to item j with 1 standing for correct and 0 incorrect, θ_i is test-taker i's ability that theoretically spans from negative infinity to positive infinity on a continuum called θ scale, a_j is item j's discrimination power, b_j is item j''s difficulty, and c_j is item j's lower asymptote, a.k.a. pseudo-guessing (Hambleton & Swaminathan, 1984). The constant of 1.7 is added to equation to make the item characteristic curve (ICC) in this logistic model resemble the ICC in the normal ogive model. If c-parameters are fixed at zero, the 3PL model would be reduced to the 2PL model. If a-parameters are further fixed at unity, the 2PL model would be reduced to the 1PL model. If the constant

1.7 is further set to unity, the 1PL model would become the Rasch model. All models described above incorporate one ability-specific parameter and at least one item-specific parameter. That means, for a calibrated item where item parameters are known, "a person's θ is all we need in order to determine his probability of success on a specific item" (Lord, 1980).

If an extra dimension consistently affects the test-taker's performance on certain items or the complete test but is not included in the construct being assessed, the unidimensionality assumption of the test is very likely violated. The affecting items might consequently show "local item dependence" (LID; Yen, 1984, 1993), meaning that responses on those items are determined by not only the ability being assessed but also something not being assessed. If items were locally independent, the likelihood of a response vector is given by

$$L(\boldsymbol{U_i}|\theta_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{j=1}^{N} P(u_{ij} = 1|\theta_i, a_j, b_j, c_j)^{u_{ij}} P(u_{ij} = 0|\theta_i, a_j, b_j, c_j)^{1-u_{ij}} \quad (2.3)$$

where *N* is the number of items, $\boldsymbol{U}$ is test-taker *i*'s vector of responses, $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are the vectors of item parameters, $P(u_{ij}=1)$ is the probability of test-taker *i* getting a correct response to item *j*, and $P(u_{ij}=0)$ is the probability of test-taker *i* getting an incorrect response to item *j*. In contrast, if items were locally dependent, the above likelihood equation does not necessarily stand up.

Yen (1984) introduced the Q_3 statistic to detect LID, and it has been widely used in practice since then (Pommerich & Segall, 2008; Yen, 1984, 1993; Zenisky, Hambleton, & Sireci, 2006). The Q_3 statistic was given by the correlation between residuals remained in an item pair following the removal of the variances explained by $\hat{\theta}$s. The computation was given by

$$d_{ij} = u_{ij} - P_{ij}(\hat{\theta}_i) \tag{2.4}$$

$$Q_{3ij} = r_{d_j d_k} \tag{2.5}$$

where d_{ij} is the test-taker i's residual on item j, and $P_{ij}\left(\hat{\theta}_i\right)$ is the expected probability of a correct response, or expected raw score (ERS). When the LID is absent between two items, the Q_3 statistic would be a small negative number from the normal distribution with $\mu = 0$ and $\sigma^2 = 1/(N-3)$. When the LID is present but is ignored in analysis, however, it would mislead analysts to overestimate test reliability and item information function (Wainer & Thissen, 1996; Yen, 1984, 1993; Zenisky et al., 2006). Occasional LIDs among items sharing the same question prompt could be treated by combining those items to a polytomously scored item and analyzing that polytomous item using the testlet model (Sireci, Thissen, & Wainer, 1991; Wainer, Bradlow, & Du, 2002; Wainer & Kiely, 1987; Wainer & Thissen, 1996). Consistent LIDs throughout the test, on the other hand, might imply the presence of multidimensionality and thereby invite the applications of MIRT techniques (Reckase, 1985, 1997, 2009; Reckase & McKinley, 1991).

The stringent definition of unidimensionality is difficult, if not impossible, to follow in reality. However, a minor violation to the unidimensionality assumption would not necessarily disqualify the UIRT analyses from being validly used. A less stringent definition of the unidimensionality was developed to determine whether the test is “essentially unidimensional” and whether the UIRT analysis is appropriate for the testing data (Nandakumar & Stout, 1993; Stout, 1987, 1990). The statistical test of the essential unidimensionality involved a nonparametric t-test comparison between two sets of items, one containing items identified to be most likely unidimensional and the other containing remaining items in the test that might be at odds with the unidimensionality presented in the first set (Nandakumar, 2005; Nandakumar & Stout, 1993; Stout, 1987, 1990, 2005). A statistical significant would suggest the presence of multidimensionality; otherwise, the test is essentially unidimensional.

Unidimensional Dual-purpose Test

In practice, test developers intentionally select items showing empirical evidence of satisfactory characteristics on the statistically most significant dimension in the field test in order to construct a unidimensional test with minimal dimensionality. While this might neglect information not reflected on the measurement dimension and lose some estimation efficiency, it ensures the test is internally consistent and able to produce reliable total score that is easy to interpret and use. Although items in a unidimensional test are deemed to measure the same construct, they can still be classified into different content domains as if measuring different aspects of the construct. From the validation perspective, the “multi-aspect” assessment of the construct is necessary, as it aligns the content representation of the test with the scope of the construct so as to avoid the construct underrepresentation error and construct irrelevant variance (Kane, 2006; Messick, 1995). Subtests based on content domains at times tempts practitioners to derive subscores from content-based subtests and assume they carry diagnostic information with respect to domain-specific abilities that could supplement the unidimensional total score obtained in the complete test. This results in a unidimensional dual-purpose test.

A well-developed unidimensional dual-purpose test is expected to yield reliable and accurate total score, because the test is essentially fine-tuned for high-quality total score. However, subscores might exhibit some problems: inaccurate, unreliable, and providing diagnostically useless information (Luecht et al., 2006; Wainer et al., 2001). For instance, Haberman (2008) developed a CTT-based method of testing whether subscores provide added-value to the total score, in which the proportional reduction of mean square error (PRMSE) of the true subscore that was estimated with observed subscore, denoted as $PRMSE_S$, was compared to the PRMSE of the true subscore that was estimated with observed total score, denoted as $PRMSE_T$. If *PRMSEs* was larger than $PRMSE_T$, then the observed subscore was considered as a more accurate estimate of the true subscore than the observed total score, and hence, provided additional diagnostic information to the observed total score. In this sense, this subscore should be reported. If $PRMSE_S$ was smaller than $PRMSE_T$ otherwise, the observed subscore was essentially useless.

Using this criterion, Haberman (2008) found that both SAT and Praxis failed to produce subscores with added value. The author further argued that subscores had to be highly reliable and somehow distinct from the total score in order to carry added value. Additional empirical studies were conducted to examine a wider array of operational unidimensional tests using this criterion, finding that most subscores were diagnostically useless (Puhan et al., 2010; Sinharay, 2010; Sinharay et al., 2007). Sinharay (2010) argued that subtests of a unidimensional test needed to comprise of at least 20 items in order to “have any hope of having added value”.

Alternative subscoring methods that exploit collateral, or ancillary, information across subtests of the complete test were proposed in hopes of deriving more accurate estimates of subscores. The collateral information refers to the information that has potential of being utilized to improve estimation but is overlooked by the traditional scoring method, such as demographics, educational background, etc. (Mislevy & Sheehan, 1989). If the collateral information of interest already exists within the test, it is termed as the in-test collateral information; conversely, if it requires additional data collection process, it is then referred to as the out-of-test collateral information (de la Torre et al., 2011). All of the alternative subscoring methods to be introduced in the following paragraphs exploit the in-test collateral information only.

One of the early attempts was Yen’s (1987) objective performance index (OPI), which used regular IRT estimates of the complete test as “prior information” to adjust subscores. The adjusted subscores supposedly have less error than raw subscores. Per Yen, the computation of OPIs took following steps. First, the complete test was analyzed as a whole to obtain item and ability parameters. Second, parameters were used to compute the expected average observe score T_i:

$$\hat{T}_i = \frac{1}{n}\sum_{j=1}^{n}\left(c_j + \frac{1-c_j}{1+\exp[-1.7a_j(\hat{\theta}_i-b_j)]}\right) \tag{2.6}$$

and the chi-square statistic *Q*:

$$Q = \sum_{i=1}^{N} \frac{(x_i/n - \hat{T}_i)^2/n}{\hat{T}_i(1-\hat{T}_i)} \tag{2.7}$$

Third, if $Q < \chi^2 (N, .10)$, the final OPI estimate was given by

$$\hat{T}_i = \frac{\hat{T}_i n^*{}_i + x_i}{n^*{}_i + n_i} \tag{2.8}$$

where

$$n_i^* = \frac{\mu(\hat{T}_i|\theta)[1-\mu(\hat{T}_i|\theta)]}{\sigma^2(\hat{T}_i|\theta)} \tag{2.9}$$

Otherwise, if $Q > \chi^2 (N, .10)$, the OPI estimate was given by

$$\hat{T}_i = \frac{x_i}{n_i} \tag{2.10}$$

Wainer et al. (2001) introduced the score augmentation method on the basis of Kelley's true score regression and empirical Bayes theorem, which used observed covariance and reliabilities to regress observed subscores toward true subscores. Like Yen's OPI, the resulting subscores supposedly exhibit higher reliability and less error than raw subscores. Kelley's true score regression was given by

$$\hat{\tau} = \rho x + (1-\rho)\mu \tag{2.11}$$

where $\hat{\tau}$ is the estimated true score, ρ is the observed reliability, x is the observed score, and μ is the mean score. The regressed score τ was a more reliable but biased estimate of the true score than raw observed score x. Rearranging terms, the equation was identical to:

$$\hat{\tau} = x_{.} + r(x - x_{.}) \tag{2.12}$$

where ρ is substituted by a sample estimate of reliability r, and μ is substituted by a sample estimate of mean x.. Extending the equation to the multivariate scenario, subscores could be augmented as follows:

$$\hat{\tau} = \boldsymbol{X}_{.} + \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{X}_{.}) \tag{2.13}$$

where the matrix $\boldsymbol{B}$ was given by

$$\boldsymbol{B} = \boldsymbol{S}^{\boldsymbol{True}}(\boldsymbol{S}^{\boldsymbol{Obs}})^{-1} \tag{2.14}$$

where $\boldsymbol{S}^{\boldsymbol{Obs}}$ and $\boldsymbol{S}^{\boldsymbol{True}}$ are the variance-covariance matrices of observed subscores and true subscores respectively. The unobserved $\boldsymbol{S}^{\boldsymbol{True}}$ matrix could be approximated using and $\boldsymbol{S}^{\boldsymbol{Obs}}$ and observed subscore reliabilities $\rho_{nn'}$. Thus, the estimation of $\boldsymbol{S}^{\boldsymbol{True}}$ is given by

$$\begin{bmatrix} s_{11}^{obs}\rho_{11'} & \cdots & s_{n1}^{obs} \\ \vdots & \ddots & \vdots \\ s_{n1}^{obs} & \cdots & s_{nn}^{obs}\rho_{nn'} \end{bmatrix}$$

Empirical evidence suggested that score augmentation was more effective for multidimensional tests than unidimensional tests (Wainer et al., 2001). For example, when subscores of the 1994 American Production and Inventory Control Society (APICS) certification examination were augmented, little advantage was gained from score augmentation, because information was barely borrowed from that unidimensional test. On the contrary, for the performance part of the North Carolina Test of Computer skills, which was a multidimensional test by design, the augmented subscores showed appreciable higher reliabilities and less error than raw subscores.

In the situation where the domain-specific abilities assessed in subtests presumably slightly deviate from the overall ability assessed in the complete test, it was sensible to apply MIRT to model these dimensional deviations and collect information across dimensions (Luecht, 1996; Reckase, Ackerman, & Carlson, 1988; Segall, 1996, 2010). For subscoring purposes, de la Torre and Patz (2005) described a MIRT-based hierarchical Bayesian subscoring method, in which several correlated but independently assessed abilities in a test battery were simultaneously estimated using a Markov Chain Monte Carlo (MCMC) estimation algorithm due to the computational complexity of the problem. Correlational collateral information was consequently borrowed across subtests to improve the subscore estimations. Results of a simulation study confirmed that compared with multiple independent unidimensional estimations, the simultaneous multidimensional estimation showed more reliable and accurate estimation results, especially for tests with highly correlated abilities.

More recently, de la Torre and Song (2009) introduced the high-order IRT (HO-IRT) subscoring method, in which secondary domain-specific abilities that determine the test-taker’s subtest performances are supposedly derived from a higher-order ability, or the primary/overall ability. While the primary ability score was deemed reflecting the test-taker’s total score on the complete test, the secondary ability scores obtained from subtests were deemed the test-taker’s subscores for corresponding subtests. The authors named this hierarchical structure as “multi-unidimensionality” in their study, suggesting that it was unidimensional on the test level but multidimensional on the subtest level. This model distinguishes itself from the bi-factor model, which would be described in later sections, in that the test-taker’s performance on an item is solely determined by secondary abilities, rather than both the primary ability and domain-specific abilities as in the bi-factor model. Results of a simulation study implied that in general, HO-IRT estimates of total scores were rather comparable to UIRT estimates, yet slightly less biased and more efficient than UIRT estimates, especially when the test was multidimensional.

In a study comparing all of these subscoring methods described above, they were found to produce very comparable subscores, except for Yen’s OPI (de la Torre et al., 2011). As the authors explained, the augmentation, MIRT, and HO-IRT subscoring methods all exploited the correlational collateral information across dimension. As a result, they were expected to deliver similar results. In addition, these alternative subscoring methods generally showed their advantages over traditional methods when dimensional correlations were high enough to allow collateral information to be borrowed. The test length did not present a significant effect, meaning longer subtests did not necessarily result in better subscores. When subtests were so long that they were already highly reliable, there was no necessity to employ any alternative subscoring technique.

Multidimensionality

The ability that a test attempts to measure or actually measures is at times multidimensional in reality, constituted of multiple correlated or uncorrelated further specific abilities. In this case, the conventional UIRT models might fall short in describing the sophisticated interaction between test-takers and testing items and provided biased and less efficient score estimations (Reckase, 2009). Take a language proficiency test for example, which attempts to measure the test-taker's abilities in four modalities by design: listening, reading, speaking, and writing. These four abilities are all related to language proficiency, but none can fully represent it. A high proficiency in one modality is not automatically related high proficiency on other modality. Therefore, the test-taker's language proficiency cannot be accurately described unless his or her abilities in all four modalities are considered. Take another math achievement test for example. The math test is designed to assess the math ability, but might unintentionally involve the reading ability. It is possible that some test-taker's incorrect responses are caused by the misunderstanding of question prompts due to a lower reading ability, instead of lower math ability. Accounting for both math and reading abilities using the MIRT model supposedly isolates the effects of the math and reading abilities, and thus, provides a more accurate estimation of the math ability than estimating it along using the UIRT model.

In general, there are two classes of multidimensional models: compensatory and noncompensatory models. The noncompensatory model theorists assumed that dimensional abilities are isolated between dimensions and a correct response to an item in this case required the test-taker to master all abilities (Sympson, 1978). Thus, the probability of getting a correct response to an item is given by

$$P\left(u_{ij} = 1 \middle| \boldsymbol{\theta_i}, \boldsymbol{a_j}, \boldsymbol{b_j}, c_j\right) = c_j + \left(1 - c_j\right)\left[\prod_{k=1}^{m} \frac{1}{1+\exp\left[-1.7a_{jk}(\theta_{ik}-b_{jk})\right]}\right] \qquad (2.15)$$

where θ_i is test-taker *i*'s *m*-dimensional vector of abilities, a_j and b_j are item *j*’s *m*-dimensional vectors of discrimination and difficulty parameters, θ_{ik} is test-taker *i*'s ability in dimension *k*, and a_{jk} and b_{jk} are item *j*’s parameters in dimension *k*.

In contrast, the compensatory model theorists assumed that deficiency in one dimension could be compensated by sufficiency in other dimensions, and the test-taker’s “total ability” that accounted for all abilities determined his or her probability of answering an item correctly (McDonald, 1985). The probability of answering an item correctly in the compensatory multidimensional normal ogive model is given by

$$P\left(u_{ij} = 1 \middle| \boldsymbol{\theta_i}, \boldsymbol{a_j}, \boldsymbol{b_j}, c_j\right) = c_j + \left(1 - c_j\right)N\left[\sum_{k=1}^{m} a_{jk}\left(\theta_{ik} - b_{jk}\right)\right] \qquad (2.16)$$

where *N(.)* is the normal distribution function. Because of the following equation

$$\sum_{k=1}^{m} a_{jk}\left(\theta_{ik} - b_{jk}\right) = \sum_{k=1}^{m} a_{jk}\theta_{ik} + \sum_{k=1}^{m} a_{jk}b_{jk} \qquad (2.17)$$

in which the second summation term on the right only involves item parameters, the multidimensional normal ogive model with reduced number of parameters is given by

$$P\left(u_{ij} = 1 \middle| \boldsymbol{\theta_i}, \boldsymbol{a_j}, d_j, c_j\right) = c_j + \left(1 - c_j\right)N\left[\sum_{k=1}^{m} a_{jk}\theta_{ik} + d_j\right] \qquad (2.18)$$

Accordingly, the logistic form of this model, which is often called the multidimensional 3PL (M3PL) model, is given by

$$P\left(u_{ij} = 1\middle|\boldsymbol{\theta_i}, \boldsymbol{a_j}, d_j, c_j\right) = c_j + \frac{1-c_j}{1+\exp\left[-1.7\left(\sum a_{jk}\theta_{ik}+d_j\right)\right]} \tag{2.19}$$

If *c*-parameters are fixed at zero, this model is reduced to the multidimensional 2PL (M2PL) model. Based on the M2PL model, some useful statistics were developed to describe item characteristics as in the UIRT, such as discrimination, difficulty, and information function (Reckase, 1985; Reckase & McKinley, 1991). By definition, the multidimensional discrimination (MDISC) is the item’s discrimination power in the direction with maximum information, and it is given by

$$MDISC = \sqrt{\sum_{k=1}^{m} {a_{jk}}^2} \tag{2.20}$$

Similarly, the multidimensional item difficulty (MDIFF) was defined as the item’s difficulty in the direction with maximum information, and it is given by

$$MDIFF = \frac{-d_j}{\sqrt{\sum_{k=1}^{m} {a_{jk}}^2}} \tag{2.21}$$

The geometrical angle between item *j*’s dimension *k* and the dimension with maximum information is given by

$$\cos \alpha_{jk} = \frac{a_{jk}}{\sqrt{\sum_{k=1}^{m} {a_{jk}}^2}} \tag{2.22}$$

The item's multidimensional information function (MINF) on direction α is given by

$$MINF = I_{\alpha}(\theta) = P_i(\theta)Q_i(\theta)(\textstyle\sum_{k=1}^{m} a_{jk} \cos \alpha_k)^2 \qquad (2.23)$$

A special case of importance is the bi-factor model, in which the multidimensionality structure consists of a primary ability dimension and *m* specific ability dimensions (Gibbons & Hedeker, 1992). The bi-factor model imposes the constraint on the item characteristics loading pattern as such that each item has nonzero loadings on the primary dimension and only one specific dimension as follows:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & \alpha_{32} & 0 \\ \alpha_{41} & 0 & \alpha_{43} \\ \alpha_{51} & 0 & \alpha_{53} \\ \alpha_{61} & 0 & \alpha_{63} \end{bmatrix}$$

where there are six items and two specific dimensions, three items per specific dimension. Under some circumstances, the abilities in specific dimensions can be understood as the strengths and weakness. Take the four-modality language proficiency for example. While the primary ability corresponds to the overall language proficiency, a positive θ score in the reading ability would increase the test-taker's probabilities of getting correct responses to all reading items and a negative θ score in the writing ability would decrease the test-taker's probabilities of getting correct responses to all writing items. The bi-factor model gives each test-taker a primary score as well as *m* specific scores. Such bi-factor constraint leads to a simpler loading pattern than regular M2PL and M3PL models, resulting in a remarkably simpler estimation.

The MIRT models were historically calibrated with factor analytical approaches (Bock, Gibbons, & Muraki, 1988; McDonald, 1982). Two widely used computer programs of MIRT calibration were NOHARM (Fraser, 1993) and TESTFACT (Wilson, Wood, & Gibbons, 1991). While both computer programs allow exploratory factor analysis (EFA), only NOHARM allows general confirmatory factor analysis (CFA). Empirical results were inconsistent with respect to the calibration performance, and no definitive evidence was found to favor one program over the other (Reckase, 2009). However, general studies regarding the factor analytical approach suggested that factor analysis was vulnerable to indeterminacies brought forth by subjective, or arbitrary, decisions on rotation methods, the number of factors, etc. (Luecht et al., 2006; McLeod, Swygert, & Thissen, 2001; Swygert, McLeod, & Thissen, 2001).

Alternative to the factor analytical approach, Luecht and Miller (1992) proposed a two-stage mixed approach to calibrating MIRT models. That is, a factor analysis and a hierarchical cluster analysis were conducted in Stage 1 to classify items into independent clusters, with each cluster representing a unidimensional subtest. The UIRT calibration was subsequently conducted to each individual subtest in Stage 2 to obtain item parameters. This two-stage approach was found to produce high level of accuracy and stability in parameter estimation and better representation of the multidimensionality structure underlying responses (Luecht & Miller, 1992).

The relationship between the UIRT and MIRT was an intriguing topic that drew much attention. Reckase et al. (1988) argued that “conceptually, any item that can be described by the M2PL model is unidimensional in that it is equivalent to an item described by a unidimensional model with the ability scale equal to a weighted composite of the elements of the θ-vector.” In other words, if multidimensional items measure the same weighted composite of multidimensional abilities throughout the test, they would eventually constitute a unidimensional test. The other way around, a unidimensional test can be thought of assessing a consistent composition of multiple specific abilities. In contrast, if the composition of specific abilities varies within the test, the test might be regarded as having idiosyncratic multidimensionality or principled multidimensionality, depending on the degree to which items are heterogeneous (Luecht & Miller, 1992; Luecht et al., 2006). When dimensions are highly correlated, the unidimensional calibration might obtain adequate estimation of item and ability parameters (Drasgow & Parsons, 1983). However, when dimensions are not highly correlated, the relationship between unidimensional estimates of multidimensional data and true multidimensional parameters could not be described by any simple linear functions. For instance, Ansley and Forsyth (1985) and Way, Ansley and Forsyth (1988) found out that for data generated with the compensatory multidimensional model, the unidimensional estimate $\hat{a}$ approximated the sum of $a_1$and a_2, and $\hat{b}$ approximated the mean of $b_1$and b_2. For data generated with the noncompensatory model, $\hat{a}$ approximated the mean of $a_1$and a_2, and $\hat{b}$ was an overestimate of $b_1$and b_2. It is reasonable to expect the relationship would be further complicated as the multidimensionality of the test increases.

Multidimensional Dual-purpose Test

In the context of the present study, it is imperative to distinguish two types of multidimensionality: the random and fixed multidimensionality (Wainer & Thissen, 1996). While the random multidimensionality refers to the presence of unexpected extraneous dimensionality in the test, which brings in error to the estimation, the fixed multidimensionality is defined as multidimensionality by design, which intends to reduce error in estimation by accounting for more sources of variances in test performances (Luecht, 1996; Segall, 1996, 2010). Thus, only the fixed multidimensionality is desirable in terms of reporting useful subscores. The fixed multidimensionality can be realized in two ways in test design: the idiosyncratic multidimensional design and the principled multidimensional design. The idiosyncratic multidimensional design is defined in this study as the design allowing factorially complex-structured items (i.e. item having nonzero loadings on multiple dimensions), as long as the marginal item characteristics agree with the test plan. Conversely, the principled multidimensional design is defined as the design allowing only factorially simple-structured items (i.e. item having nonzero loading on one dimension) in the test, which leads to unidimensional subtests with prescribed relationship between subtests (Luecht et al., 2006). This stringent constraint imposed on the principled multidimensional design arguably leads to scales with higher statistical stability, psychometric quality, and interpretability (Luecht et al., 2006), and it is in compliance with recent advancements of test design theory—evidence centered design (Hendrickson, Huff, & Luecht, 2010; Huff, Steinberg, & Matts, 2010; Luecht et al., 2006; Mislevy, 1994, 1996).

On the basis of the idiosyncratic multidimensional design, the conventional MIRT model exploits both idiosyncratic dimensional information that is absent or uncollected in the composite dimension and cross-dimensional collateral information. Therefore, it supposedly improves the estimations of domain-specific abilities (Luecht, 1996; Reckase, 2009; Segall, 1996, 2010). The implementation of MIRT at times was confronted by the calibration problem. In the scenario where analysts attempt to detect the presence of any unintentional dimension(s), an exploratory factor analytical calibration might be conducted to find a solution on the basis of statistical associations and differentiations among items. The subjective decisions are then necessary for determining the number of factors and the rotational method. This purely data-based analysis might be too sensitive to unsystematic errors in the data to yield consistent and stable calibration results across administrations of the same test (Luecht et al., 2006; McLeod et al., 2001; Swygert et al., 2001). Conversely, if analysts acquire strong pre-knowledge regarding the item-factor loading structure, the confirmatory factor analytical calibration might be more appropriate than the EFA calibration, because the CFA calibration tends to yield consistent results that are compatible results with the test plan. It should be remembered that in order for the CFA calibration to be viable, “for each trait there are at least two items measuring it that are factorially simple” (McDonald, 2000).

Compared to the idiosyncratic multidimensional design, the principled multidimensional design is arguably easier to achieve and manage from the item development perspective. That is, a test of the principled multidimensionality can be thought of be composed of several unidimensional subtests, or independent item clusters, and the goal of test development is, as in a unidimensional test, to develop items primarily measuring sole one ability for each subtest. Although the exploratory and confirmatory factor analytical calibrations might still be applicable, Luecht et al. (2006) argued that separate UIRT calibrations, one per each essentially unidimensional subtest, showed minimal technical inferiority to the sophisticated and computationally intensive MIRT calibration.

Another difficulty related the multidimensional dual-purpose test is the derivation of the total score. While the MIRT model only provides the estimation of subscores, the total score needs to be accurately derived with alternative methods to represent the test-taker's overall ability with an easily interpreted score for the ranking purpose of the test. A straightforward method is to conduct a unidimensional analysis of the test in addition to the multidimensional analysis (Brandt, 2008, 2010). As Brandt argued, although this practice was psychometrically sound, it was difficult to explain to the public the interpretations of different parameters for the same items. As described in preceding section, the HO-IRT might also be plausible in resolving this problem; however, it imposed a relatively novel and less-studied model. Instead, some score-aggregation methods were proposed to combine domain-specific scores to a composite total score. The simplest method is to estimate the total score by summing raw subscores or averaging θ subscores. This method is logically straightforward and computationally convenient, but it results in an inefficient estimate with larger variance and more error than any of the subscores.

Graybill and Deal (1959) introduced a method that combined estimates of the population mean from two independent samples and resulted in a uniformly better estimate than either individual sample estimate:

$$\hat{\mu} = (n_1 {s_2}^2 x + n_2 {s_1}^2 y)/(n_1 {s_2}^2 + n_2 {s_1}^2) \tag{2.24}$$

where *x* is distributed as a normal variable with mean μ and variance σ_1^2/n_1, *y* is also distributed as a normal variable with mean μ and variance σ_2^2/n_2, and s_1 and s_2 are sample standard deviations. The estimate $\hat{\mu}$ obtained in this approach would have less or equal variance to both σ_1^2 and σ_2^2, meaning that it is a uniformly better estimate. Assuming subscores are independent estimates of the overall ability, this approach might be employed to derive a weighted composite total score that is better than individual subscores.

Another method was introduced by Longford (1997) to specifically address the issue of combining estimates of test scores. Supposing arbitrary weights were assigned prior to the data collection, these weights were then adjusted using the empirical post-administration data so as to produce a composite score showing minimal (conditional) mean square error (MSE):

$$\boldsymbol{v}^* = \boldsymbol{S}_{Obs}^{-1} \boldsymbol{S}_{True} \boldsymbol{w} \tag{2.25}$$

where $\boldsymbol{S_{Obs}}$ and $\boldsymbol{S_{True}}$ are the variance-covariance matrices of the observed and true subscores. Subsequently, the composite total score X_v was given by

$$X_v = \boldsymbol{\mu}'\boldsymbol{S}_{Obs}^{-1}\boldsymbol{S}_{Error}\boldsymbol{w} + \boldsymbol{X}'\boldsymbol{S}_{Obs}^{-1}\boldsymbol{S}_{True}\boldsymbol{w} \tag{2.26}$$

where μ is the *m*-dimensional vector of score means.

Lastly, a factor-analytic method was introduce by Luecht et al. (2006) to approximate the composite total score using observed subscores. This method obtained the total score that would explain maximal variance among subscores. To do so, a factor analysis was conducted on standardized subscores, obtaining the largest eigenvalue *S* and the corresponding eigenvector $\boldsymbol{A}$, which were then used to approximate score-aggregating weights $\boldsymbol{W}$ as follows:

$$\boldsymbol{P}' = \sqrt{S}\boldsymbol{A}' \tag{2.27}$$

$$\boldsymbol{W}' = \boldsymbol{P}'(\boldsymbol{P}'\boldsymbol{P})^{-1} \tag{2.28}$$

And, the composite total score θ^* was given by

$$\theta_i^* = \theta_i\boldsymbol{W} \tag{2.29}$$

Purpose and Questions of Research

Embedding both the ranking and diagnosis purposes in one assessment instrument would be great advancement in this field. Our understandings of the optimal design and analysis of such dual-purpose test, however, are dwarfed by the growing needs for it in practice. Potential psychometric challenges related to the dual-purpose testing were not fully addressed in the literature. The purpose of the present study was, thus, to provide a systematic comparison of a variety of possible designing and analyzing paradigms for the dual-purpose test in conditions with varying test length and dimensionality structure of true abilities.

The above considerations led to four main research questions listed below:

1. In which design(s) can the total score be most accurately and reliably estimated?

    1.1. Is this effect consistent over various dimensionality structures of true abilities?

    1.2 Is this effect consistent over various test lengths?

    1.3. Is this effect consistent over various calibration approaches?

2. In which design(s) can subscores be most accurately and reliably estimated?

    2.1. Is this effect consistent over various dimensionality structures of true abilities?

    2.2 Is this effect consistent over various test lengths?

    2.3. Is this effect consistent over various calibration approaches?

3. In which design(s) can subscores provide added value to the total score?

    3.1. Is this effect consistent over various dimensionality structures of true abilities?

    3.2 Is this effect consistent over various test lengths?

    3.3. Is this effect consistent over various calibration approaches?

4. Do alternative scoring methods, inclusive of both alternative subscoring methods and score-aggregating methods, improve the accuracy and reliability of raw estimated scores?

    4.1. Is this effect consistent over various test designs?

    4.2. Is this effect consistent over various dimensionality structures of true abilities?

    4.3. Is this effect consistent over various test lengths?

    4.4. Is this effect consistent over various calibration approaches?

CHAPTER III

METHODS

Context of Study

A hypothetical math test was introduced as a concrete example to illustrate the methodology. This math test was designed to measure mathematical skills and knowledge that students learned in schools, consisting of questions coming from four content domains: arithmetic operations (ARI), algebra (ALG), geometry (GEO), and statistics and probability (STA). These domains were denoted as Domain 1 to 4 in the remainder of the article. Whether this multi-domain test was unidimensional or multidimensional depended on the actual test designs, which was manipulated in this study. If the test was unidimensional, domains represented various components of the overall math ability. If the test was multidimensional, domains essentially represented various dimensions folded in the overall math ability. Moreover, since domains were not likely to display equivalent difficulties in reality (i.e., some domain being easier while others being more challenging), item difficulties were presumably compounded with domains as such that the ARI items were easiest, having average item difficulty equal to -.5, followed by the ALG ($\bar{b} = -.2$), GEO ($\bar{b} = .2$), and STA items ($\bar{b} = .5$).

With respect to scores, there were two intended functions, or utilities. The primary function was to report a total score for each student, indicating the student’s overall math ability. This total score would be utilized to assist educational institutions in ranking students and making critical decisions accordingly. The secondary function was to report diagnostic subscores, assessing various components/dimensions of the overall math ability, namely, the abilities of arithmetic operations, algebra, geometry, and statistics and probability respectively. Not only did subscores offer educational institutions finer-grained information to further rank students, but also directed the student's focus onto identified weak areas or abilities. Therefore, the test met the definition of the dual-purpose test in this study, and was required to report both reliable and accurate total score and domain-specific subscores.

Conditions of Study

Three factors were manipulated in this study, namely the test design, test length, and calibration approach, resulting in 7×3×3 cross-factor experimental conditions. The test design was manipulated with seven levels: one unidimensional (UNI) design, three idiosyncratic multidimensional (IM) designs with dimensions correlated at ρ=.3, .6, and .9, and three principled multidimensional (PM) designs with dimensions correlated at ρ=.3, .6, and .9. The UNI design was to mirror the practice as such that items were meticulously selected for the inclusion in the test based on statistical estimates of their characteristics on the single dimension (the "reference composite" dimension, if it were a "true" multidimensional test) obtained from the field test in order to construct a psychometrically sound and robust unidimensional test. It was also reasonable to assemble test with multidimensional items, maximizing the diagnostic functionality of the test. The multidimensionality could be idiosyncratic, in which each item provided information on one major dimension as well as several minor dimensions, or principled, in which each item was psychometrically effective on one predefined dimension. Specifically, the IM designs imitated the practice of constructing a conventional multidimensional with idiosyncratic information, whereas the PM designs imitated the practice of imposing principled information on test designs ((Luecht et al., 2006; Mislevy, 1994, 1996).

The test length was manipulated with three levels: 5 items per domain (N=5), 10 items per domain (N=10), and 20 items per domain (N=10). Since the test was composed of four domains, these three levels corresponded to the total test length of 20, 40, and 80 respectively. The 20-item test (5 items per domain) represented the short test-length condition, which was unlikely for any high-stake tests but might be plausible for the low-stake formative classroom tests. In contrast, the 80-item test (20 items per domain) represented the long test-length condition. The length of 80 items might still be short for high-stake tests, but the effect of test-length should manifest itself through the progression from 5 items per domain to 20 items per domain.

Crossing these two factors resulted in 7 × 3 “data generation conditions”, in each of which a data set of responses with 5,000 examinees and corresponding number of items was generated. Afterwards, the data set was calibrated with three approaches: the concurrent unidimensional (CU) calibration, the confirmatory multidimensional (CM) calibration, and the separate unidimensional (SU) calibration. The CU approach was the conventional unidimensional calibration approach, in which all items were simultaneously calibrated as a whole and assigned with parameters indicating their characteristics on the calibrated dimension. The CM approach defined a simple-structured confirmatory factor loading pattern for calibration as such that the first *N* items in an *N*-item-per-domain test assessed the first dimension, the second *N* items the second dimension, and so on. The SU approach, instead, calibrated each subtest with a unidimensional model individually.

In summary, crossing all factors resulted in 63 experimental conditions. Item parameters, ability parameters, and responses were generated according to the test design in each of 21 data generation conditions. The sample size was held constant at 5,000 in all data generation condition, which is supposedly sufficient to produce stable estimation results. Also, 30 replications were conducted to reduce sampling error incurred in data generation.

Data Generation

Item Parameters

The data generation began with the item generation for the IM design. For the IM designs with *N* items per domain, a *4N×4* matrix of *a*-parameters was generated from the lognormal distribution with the following mean vector and variance-covariance matrix on the log scale:

$$\mu = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\sigma^2 = \begin{bmatrix} .64 & & & \\ 0 & .64 & & \\ 0 & 0 & .64 & \\ 0 & 0 & 0 & .64 \end{bmatrix}$$

The *a*-parameters outside of the interval [0.0, 2.0] were regenerated until all *a*-parameters stayed within the bounds, which avoided unrealistic parameters. The resulting *a*-parameters were expected to be tightly distributed around 1.0. The order of the *a*-parameter in each row was then rearranged as such the greatest *a*-parameter was placed under the item's major assessment domain. For instance, the ARI items were generated to have largest *a*-parameters in the ARI domain.

For the PM design with *N* items per domain, the same *a*-parameters in the corresponding IM design were used to compute the pseudo-composite *a*-parameters as follows:

$$a_j = \sqrt{\sum_{k=1}^{m} a_{jk}^2} \tag{3.1}$$

This pseudo-composite *a*-parameter was also an item's MDISC, approximating the maximum information it had in the aggregate. Each item's pseudo-composite *a*-parameter was then assigned to the item's major assessment domain, leaving other three minor domains zeros. This ensured the equality of measurement information across test designs, as well as the comparability of performances of different test designs. Similarly, the pseudo-composite *a*-parameters were also assigned to items in the UNI design as their *a*-parameters, ensuring the equality of measurement information and comparability of various test designs.

In all designs, the difficulty parameters were generated from the normal distributions with the same variance of .8 but differential means for different domains in order to reflect differential hypothetical domain difficulties. The generation of difficulty parameters was bounded by the lower bound of -3.0 and upper bound of 3.0, in an attempt to avoid unrealistic difficulty parameters. The lower-asymptote *c*-parameters were generated with fixed value at .10.

Table 1. An Example of Item Generation Results for the IM Design with *N*=5

a_1	a_2	a_3	a_4	d	c	Domain
0.65	0.16	0.44	0.44	1.11	0.1	ARI
0.96	0.46	0.86	0.91	-0.02	0.1	ARI
0.83	0.38	0.21	0.33	0.55	0.1	ARI
0.63	0.15	0.40	0.49	-0.48	0.1	ARI
0.97	0.74	0.48	0.50	-1.87	0.1	ARI
0.08	**0.90**	0.34	0.21	0.68	0.1	ALG
0.15	**0.48**	0.43	0.36	0.87	0.1	ALG
0.27	**0.54**	0.49	0.48	-1.59	0.1	ALG
0.38	**0.93**	0.80	0.31	-0.79	0.1	ALG
0.19	**0.75**	0.34	0.36	-0.49	0.1	ALG
0.19	0.34	**0.36**	0.07	0.67	0.1	GEO
0.28	0.51	**0.75**	0.32	-0.69	0.1	GEO
0.45	0.11	**0.60**	0.43	0.28	0.1	GEO
0.55	0.35	**0.99**	0.14	0.41	0.1	GEO
0.37	0.34	**0.62**	0.09	0.38	0.1	GEO
0.58	0.13	0.09	**0.74**	0.46	0.1	STA
0.40	0.56	0.28	**0.73**	0.52	0.1	STA
0.30	0.63	0.21	**0.80**	-1.28	0.1	STA
0.13	0.38	0.18	**0.55**	-0.63	0.1	STA
0.16	0.19	0.34	**0.46**	-1.46	0.1	STA

Table 2. An Example of Item Generation Results for the PM Design with *N*=5

a_1	a_2	a_3	a_4	d	c	Domain
0.91	0.00	0.00	0.00	0.41	0.10	ARI
1.64	0.00	0.00	0.00	1.40	0.10	ARI
0.99	0.00	0.00	0.00	0.16	0.10	ARI
0.90	0.00	0.00	0.00	0.00	0.10	ARI
1.40	0.00	0.00	0.00	2.53	0.10	ARI
0.00	**0.99**	0.00	0.00	0.54	0.10	ALG
0.00	**0.75**	0.00	0.00	0.69	0.10	ALG
0.00	**0.92**	0.00	0.00	-0.16	0.10	ALG
0.00	**1.32**	0.00	0.00	2.00	0.10	ALG
0.00	**0.92**	0.00	0.00	0.42	0.10	ALG
0.00	0.00	**0.54**	0.00	-0.32	0.10	GEO
0.00	0.00	**1.00**	0.00	0.00	0.10	GEO
0.00	0.00	**0.87**	0.00	0.25	0.10	GEO
0.00	0.00	**1.19**	0.00	-1.15	0.10	GEO
0.00	0.00	**0.81**	0.00	-0.90	0.10	GEO
0.00	0.00	0.00	**0.96**	-0.69	0.10	STA
0.00	0.00	0.00	**1.04**	-0.33	0.10	STA
0.00	0.00	0.00	**1.08**	-2.32	0.10	STA
0.00	0.00	0.00	**0.71**	-0.65	0.10	STA
0.00	0.00	0.00	**0.62**	-0.59	0.10	STA

Ability Parameters and Scores

For each IM or PM design with ρ true ability correlation, a *5000×4* matrix of true domain-specific ability parameters was randomly from the multivariate normal distribution with the following mean vector and variance-covariance matrix:

$$\mu = [0 \quad 0 \quad 0 \quad 0],$$

$$\sigma^2 = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho & \rho & \rho & 1 \end{bmatrix}.$$

For each UNI design, 5000 unidimensional ability parameters were generated from the normal distribution with mean equal to 0.0 and standard deviation 1.0. Because of the unidimensionality, these ability parameters indicated both the examinees' overall abilities and domain-specific abilities.

The probabilities of getting a correct response were obtained by feeding item and ability parameters into the M3PL model for each IM or PM design and the 3PL model for each UNI design. These probabilities could be thought of each test-taker's expected score on each item. Therefore, the summation over a subtest gave the expected raw score (ERS) of that subtest domain, or true subscore, and the summation over the complete test gave the ERS of the test, or true total score. It was these true scores rather than true ability parameters on the θ scale that would be compared with and evaluate estimated scores in each experimental condition. This avoided complicated conversion between overall and domain-specific θ scores, for which no perfect solution was available so far.

Responses

Test-taker i's probability of answering item j correctly p_{ij} was also used to generate this test-taker's response to this item u_{ij}. Comparing with a random number r_{ij} from the uniform distribution *U(0, 1)*, the response was generated according to the following rule:

$$u_{ij} = \begin{cases} 1, if\ p_{ij} \geq\ r_{ij} \\ 0, if\ p_{ij} < r_{ij} \end{cases} \tag{3.2}$$

Calibration and Scoring

All calibrations were conducted using NOHARM (Fraser, 1993). Because the factor analytical calibration used in NOHARM required user-defined *c*-parameters, all *c*-parameters were fixed at .10 as what their true values were. The CU calibration was specified as an exploratory one-factor solution in NOHARM, yielding one *a*-parameter and one *d*-parameter (i.e., *b*-parameter with reverse sign) per item. It supposedly captured item characteristics on the statistically most significant dimension and aligned item parameters on the same dimension. The expected a posterior (EAP) estimates of the unidimensional abilities were estimated with13 quadrature points, ranging from -3.0 to 3.0 in increments of .5, from the normal distribution *N(0, 1)*. The posterior distribution was given by

$$P(\theta|\boldsymbol{u}) \propto L(\boldsymbol{u}|\theta) \times P(\theta) \tag{3.3}$$

Test-taker *i*'s final EAP estimate was given by

$$EAP = E[\theta] = \sum_{h=1}^{13} \theta_h P(\theta_h) \tag{3.4}$$

where θ_h was quadrature point *h*, and $P(\theta_h)$ was the posterior probability of quadrature point *h*. With the calibrated items and estimated abilities, the probabilities were then computed using the 3PL model. Summing the probabilities by domain resulted in estimated expected raw subscores, called raw subscores or estimated subscores, and summing the probabilities over the complete test resulted in estimated expected raw total score, called raw total score or estimated total score.

The SU calibration estimated each individual subtest separately with the exploratory one-factor solution in NOHARM. This also yielded one a-parameter and one d-parameter per item; however they were not necessarily on the same dimension. That is, parameters of an ARI item only indicated its characteristics related the ARI subtest, neither directly related to other three subtests nor to the overall test. Four EAP estimates of domain-specific abilities were estimated separately using item parameters and responses associated with each subtest, and the probabilities were computed using 3PL model. Summing the probabilities by domain resulted in raw subscores, and summing the probabilities over the complete test resulted in raw total score.

The CM calibration was specified as a confirmatory four-factor solution in NOHARM. The confirmatory loading pattern was of simple structure, in which an item had nonzero loading in its major assessment dimension but zeros in other three minor dimensions. In addition to a- and d-parameters, correlations between abilities were also estimated in NOHARM. Following the calibration, four EAP estimates of domain-specific abilities were jointly estimated using 7^4 quadrature points from the multivariate normal distribution and estimated correlations. The 7^4 quadrature points consisted of 7 quadrature points per domain, ranging from -3.0 to 3.0 in increments of 1.0. The joint posterior distribution was given by

$$P(\theta_1, \theta_2, \theta_3, \theta_4 | \boldsymbol{u}) \propto L(\boldsymbol{u} | \theta_1, \theta_2, \theta_3, \theta_4) \times P(\theta_1, \theta_2, \theta_3, \theta_4) \tag{3.5}$$

Test-taker *i*'s final EAP estimates were given by

$$EAP = E[\boldsymbol{\theta_i}] = \boldsymbol{\theta_i P}(\boldsymbol{\theta_i} | \boldsymbol{u}) \tag{3.6}$$

where θ_i was the vector of examinee *i*'s domain-specific abilities, and $P(\theta_i|u)$ was the matrix of posterior probabilities. The probabilities were then computed with the M3PL model using the item and ability parameters. Summing the probabilities by domain resulted in raw subscores, and summing the probabilities over the test resulted in raw total score.

In addition to raw total score and subscores, a set of subscores and two sets of total scores were computed using alternative socring methods. First, raw subscores were augmented using the procedure described in Wainer et al. (2001). Next, raw subscores were aggregated to create the composite total score using the methods described in Luecht et al. (2006) and Longford (1997). The former method gave a factor analytically optimal total score, denoted as the FA composite (total) score, which explained argest variance among subscores, while the latter method gave a score with least MSE, denoted as the MSE composite (total) score. The computations were documented in details in Chapter II.

Analyses

One of the most important criteria to evaluate the effectiveness of various test designs was the score accuracy, which was the extent to which estimated scores recovered their true values. The score accuracy was evaluated via Pearson's product-moment correlation and the root mean square error (RMSE). The Pearson's product-moment correlation was given by

$$r = \frac{\sigma_{\hat{X}X}}{\sigma_{\hat{X}}\sigma_X} \tag{3.7}$$

where $\sigma_{\hat{X}X}$ is the covariance of true score and estimated score, $\sigma_{\hat{X}}$ and σ_X are standard deviations of estimated score and true score. The RMSE was given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{X}_i - X_i\right)^2} \tag{3.8}$$

where $\hat{X}_i$ is test-taker *i*'s estimated score and X_i is test-taker *i*'s true score. The correlation should be considered in conjunction with the RMSE, and vice versa. High correlation and low RMSE would suggest that scores were well estimated with high level of accuracy. High correlation and high RMSE would suggest that scores might have been accurately estimated on a different scale with the true scores. Low correlation and low RMSE would suggest the restriction of range of the estimated or true scores. Low correlation and high RMSE confirmed that estimated scores were not accurate enough.

Another essential criterion to evaluate score quality was the reliability. By definition, the reliability was given by

$$\rho_{XX'} = \frac{\sigma_{True}^2}{\sigma_{Obs}^2} \tag{3.9}$$

Since true scores were generated in the simulation study, the reliability was simply squared Pearson’s correlation. By theory, the reliability should range from 0.0 to 1.0, with the higher value suggesting more reliable and less erroneous scores. As rule of thumb, a reliability of at least .85 was considered acceptable for any large-scale high-stakes examination. Otherwise, scores were contaminated with large proportion of error.

For subscores, an additional evaluative criterion was whether they provided added values to total score. Under some circumstances, subscores might be reliably and accurately estimated but failed to provide significant diagnostic information as they were expected to. In this case, subscores were still not good enough for the dual-purpose test, even though their regular psychometric qualities were satisfactory. The PRMSE criterion described in Haberman (2008), was used to determine whether subscores provided added value. The PRMSE compared the proportional reduction of mean square error of subscores that were estimated with observed subscores, denoted as $PRMSE_X$, and the PRMSE of subscores that were estimated with observed total score, denoted as $PRMSE_T$, which were given by

$$PRMSE_X = \rho^2(S_x, \tau_x) = 1 - \frac{\sigma^2(e_x)}{\sigma^2(S_x)} \tag{3.10}$$

$$PRMSE_Z = \rho^2(S_z, \tau_x) = \rho(S_z, \tau_z)\left[\frac{\rho(S_x, S_z)}{\rho(S_x, \tau_x)\rho(S_z, \tau_z)} - \frac{\sigma^2(e_x)}{\sigma(\tau_x)\sigma(S_z)}\right] \tag{3.11}$$

where X is the observed subscore, Z was the observed total score. If $PRMSE_X$ was greater, subscores provided additional information to the total score.

In addition, Q_3 was computed using estimated item and ability parameters in each experimental condition to examine the amount of unexplained covariance between items. On the basis of residual correlation, Q_3 indicated the modeling sufficiency of the calibration approach to the data, or model fit in some senses. If there was a merely insignificant correlation between residuals of an item pair, Q_3 should be a slight negative number, whose r-to-z transformation supposedly had a normal distribution with mean equal to 0 and variance 1/(N-3). Therefore, the r-to-z transformation of each Q_3 was compared with the critical value with α=.001 to determine whether it was a statistical outlier. Because it was unconcerned in this study which pair of items shared significant amount of unexplained covariance, the percentage of Q_3 outliers of all possible item pairs was used to evaluate the modeling sufficiency in each condition. The higher the percentage was, the higher the likelihood was that there was at least one significant dimension not being accounted for by the model. The Q_3 statistic expectedly provides additional supporting information to the interpretation of results.

Lastly, two widely-used model comparison statistics, the Akaike information criterion (AIC) and Bayesian information criterion (BIC), were computed to further compare calibration approaches. They assessed the goodness of fit of competing models to the data, while taking into account of model complexity at the same time. The absolute AIC/BIC value in isolation was meaningless, because they were expected to be interpreted in relative to one another. The model with lower AIC/BIC value was preferred, suggesting a higher likelihood for the model to minimize the information loss when it was used to reproduce the data. Computationally, AIC and BIC were given by

$$AIC = 2k - 2\ln(L) \tag{3.12}$$

$$BIC = k\ln(N) - 2\ln(L) \tag{3.13}$$

where *k* was the number of estimated parameters, *N* was the sample size, and *L* was the likelihood function. Both AIC and BIC were based on the negative 2 log-likelihood, but BIC incorporated a harsher penalty to overparameterization than AIC, being prone to favor the more parsimonious model.

CHAPTER IV

RESULTS

True Item Parameters and Scores

Table 3 and 4 provided descriptive statistical summaries of generated item parameters. For the UNI designs, *b*-parameters were converted to *d*-parameters by reversing the sign. For the IM and PM designs, the composite *a*-parameters, or *MDISCs*, were computed by taking the square root of the sum of squared domain-specific *a*-parameters, indicating the amount of information that a multidimensional item provided. The results showed that the means and standard deviations of *a*-parameters were very comparable across conditions (.970 for *N*=5, .986 for *N*=10, and .986 for *N*=20), which ensured the equality of measurement information across test designs. The standard deviations were also reasonably small in all conditions (.235 for *N*=5, .243 for *N*=10, and .247 for *N*=20), suggesting that the vast majority of *a*-parameters were distributed within a range bounded by .5 and 1.5. The means and standard deviations of *d*-parameters varied with conditions, but the variations were within an acceptably narrow range (means ranging from -.016 to .037, and standard deviations ranging from .788 to .945), which equated the average difficulty of the tests. The means of domain-specific *a*-parameters in the PM designs, ranging from .238 to .254, were smaller than those in the IM designs, ranging from .434 to .454, because of the simple-structure constraint imposed on the PM designs. Together, items were generated as intended.

Table 3. Means of Generated Item Parameters

	a	d	a_1	a_2	a_3	a_4
UNI						
N = 5	.970	-.008	-	-	-	-
N = 10	.986	-.016	-	-	-	-
N = 20	.986	.000	-	-	-	-
IM						
N = 5	.970	.032	.435	.448	.434	.454
N = 10	.986	.015	.447	.448	.449	.454
N = 20	.986	.037	.449	.450	.444	.452
PM						
N = 5	.970	.023	.242	.245	.238	.246
N = 10	.986	.003	.248	.239	.245	.254
N = 20	.986	.032	.248	.247	.248	.243

Table 4. Standard Deviations of Generated Item Parameters

	a	d	a_1	a_2	a_3	a_4
UNI						
N = 5	.235	.894	-	-	-	-
N = 10	.243	.861	-	-	-	-
N = 20	.247	.862	-	-	-	-
IM						
N = 5	.235	.788	.229	.235	.234	.232
N = 10	.243	.830	.241	.231	.233	.243
N = 20	.247	.811	.240	.239	.239	.236
PM						
N = 5	.235	.945	.443	.447	.435	.453
N = 10	.243	.930	.449	.440	.444	.459
N = 20	.247	.899	.449	.448	.449	.441

Table 5 and 6 presented the means and standard deviations of true scores. The means of total scores were roughly 11, 22, and 44 for the 20-item (*N*=5), 40-item (*N*=10), and 80-item (*N*=20) tests. While the means were proportional to the test length, they remained almost invariant across test designs. The standard deviations of true total scores, on the other hand, varied along with the test designs as well as test length. The score variation, on average, was largest in the IM designs but least in the PM designs. Specifically in multidimensional designs (i.e. IM and PM designs), larger score variance was observed for conditions with higher ability correlation (*ρ*=.9). Of all subscores, the means of Subscore 1 were the largest, followed by Subscore 2, 3, and 4, which reflected the differences in hypothetical domain difficulties, that is, Subtest 1 (ARI) being easier and Subtest 4 (STA) harder. For each individual subscore, it showed similar cross-condition patterns as the total score.

Presented in Table 7 were correlation coefficients between true subscores. As expected, true subscores were highly correlated at nearly 1.0 in the UNI designs, since they were driven by the same unidimensional ability. Subscores were correlated approximately at .3, .6, and .9 in the PM designs, consistent with the true ability correlations. In the IM designs, however, subscores were highly correlated at at least .918, regardless of true ability correlations. This inflation of score correlations was possibly because that each item's true ERS was a score on the local reference composite score, which considered information in dimensions instead of solely in the major dimension. Furthermore, Figure 1 presented a random example of true score distributions with10 items per domain. The example reiterated the interpretations above, and thus would not be discussed further.

Table 5. Means of True Scores

	Total Score	Subscore 1	Subscore 2	Subscore 3	Subscore 4
UNI					
N = 5	10.992	3.387	3.001	2.519	2.085
N = 10	21.873	6.498	5.868	5.109	4.398
N = 20	44.045	13.246	11.851	10.147	8.801
IM, ρ = .3					
N = 5	11.105	3.018	2.812	2.705	2.57
N = 10	22.116	5.892	5.701	5.43	5.093
N = 20	44.429	11.719	11.748	10.652	10.311
IM, ρ = .6					
N = 5	11.106	2.997	2.809	2.714	2.586
N = 10	22.026	5.838	5.662	5.418	5.108
N = 20	44.426	11.663	11.692	10.693	10.378
IM, ρ = .9					
N = 5	11.104	2.981	2.806	2.717	2.600
N = 10	22.069	5.826	5.662	5.433	5.147
N = 20	44.436	11.629	11.652	10.724	10.431
PM, ρ = .3					
N = 5	11.072	3.485	3.032	2.486	2.069
N = 10	22.018	6.848	5.887	5.019	4.264
N = 20	44.458	13.31	12.079	10.158	8.911
PM, ρ = .6					
N = 5	11.078	3.483	3.035	2.485	2.076
N = 10	22.046	6.843	5.896	5.030	4.277
N = 20	44.426	13.293	12.065	10.152	8.916
PM, ρ = .9					
N = 5	11.072	3.481	3.033	2.484	2.074
N = 10	22.032	6.841	5.887	5.025	4.279
N = 20	44.519	13.327	12.095	10.172	8.926

Table 6. Standard Deviations of True Scores

	Total Score	Subscore 1	Subscore 2	Subscore 3	Subscore 4
UNI					
N = 5	4.194	1.037	1.092	1.099	1.014
N = 10	8.490	2.126	2.161	2.179	2.089
N = 20	16.952	4.226	4.388	4.332	4.144
IM, ρ = .3					
N = 5	4.868	1.249	1.257	1.260	1.259
N = 10	9.735	2.519	2.452	2.521	2.513
N = 20	19.475	5.017	4.980	5.000	4.993
IM, ρ = .6					
N = 5	5.443	1.379	1.384	1.387	1.389
N = 10	10.856	2.771	2.702	2.772	2.764
N = 20	21.798	5.532	5.506	5.522	5.511
IM, ρ = .9					
N = 5	5.845	1.471	1.474	1.475	1.478
N = 10	11.651	2.946	2.880	2.950	2.945
N = 20	23.38	5.883	5.861	5.877	5.866
PM, ρ = .3					
N = 5	2.849	0.987	1.079	1.077	1.027
N = 10	5.787	2.071	2.120	2.189	2.089
N = 20	11.611	4.169	4.284	4.370	4.165
PM, ρ = .6					
N = 5	3.444	0.988	1.079	1.079	1.029
N = 10	7.006	2.068	2.121	2.192	2.090
N = 20	14.075	4.174	4.292	4.365	4.174
PM, ρ = .9					
N = 5	3.964	0.991	1.080	1.079	1.029
N = 10	8.060	2.070	2.120	2.191	2.095
N = 20	16.195	4.171	4.286	4.367	4.167

Table 7. Correlations between True Subscores

	N=5	N=10	N=20
UNI	.969	.979	.978
IM, ρ = .3	.918	.928	.932
IM, ρ = .6	.954	.963	.967
IM, ρ = .9	.976	.984	.988
PM, ρ = .3	.286	.288	.289
PM, ρ = .6	.573	.578	.580
PM, ρ = .9	.866	.872	.878

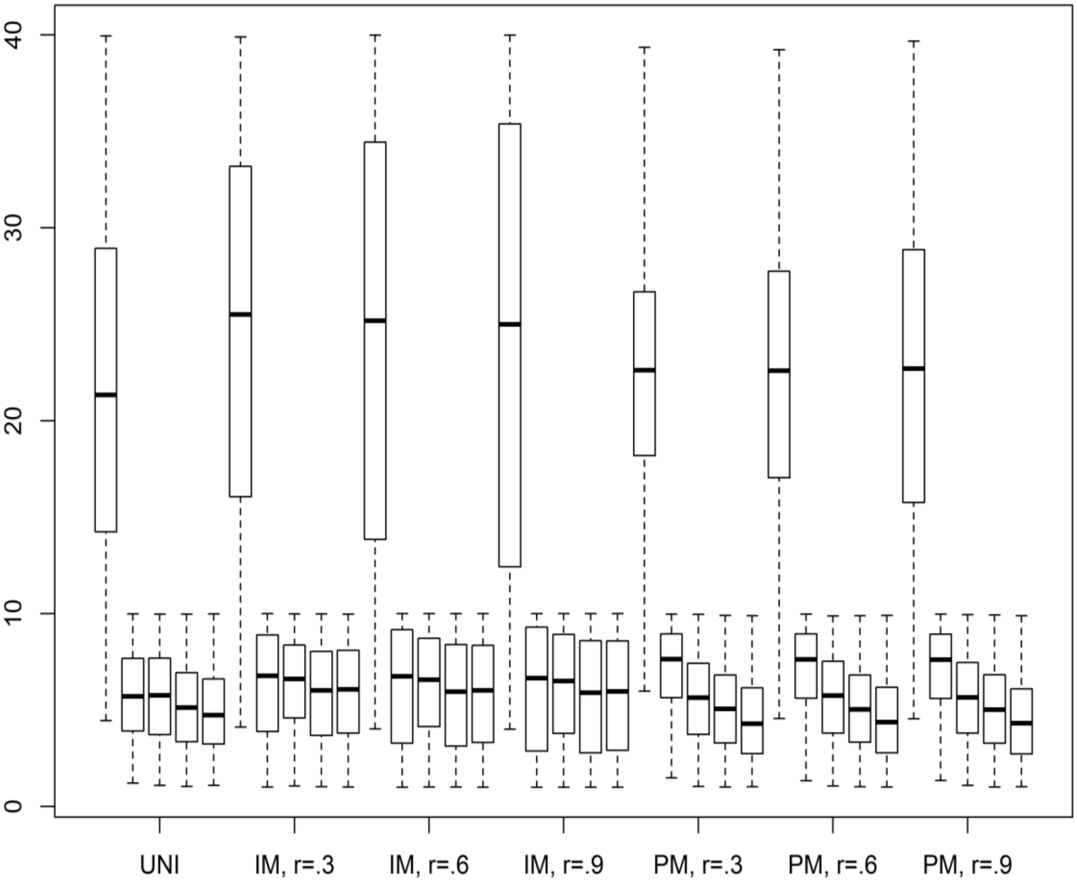

Figure 1. True Score Distributions when *N* = 10

Estimated Item Parameters

Table 8 summarized the means of item parameters when the data set was calibrated with the CU calibration approach. As it should, the CU calibration provided a good recovery of item parameters for the UNI designs, in which the means of estimated *a*- and *d*-parameters were close to the means of true parameters. However, the CU calibration overestimated *a*-parameters for the IM designs, especially for the IM designs with high true ability correlations (ρ=.9). The means of estimated *a*-parameters were all above 1.0, while the means of true *a*-parameters were slightly below 1.0. In the worst case where true abilities were correlated at .9 and test length was 20 items per domain, the mean of *a*-parameters was inflated to 2.220. On the other hand, the CU calibration underestimated *a*-parameters for the PM designs. The lower the ability correlation was, the greater the underestimation was. In the worst case where the abilities were correlated at .3 and the test length was 5 items per domain, the mean of a-parameters was deflated to .518. In general, *d*-parameters were well estimated, whose means were close to the true values in all conditions.

Table 8. Estimated Item Parameters in the CU Calibration

	N = 5		N = 10		N = 20	
	a	d	a	d	a	D
UNI	.957	-.015	.979	-.018	.976	.002
IM, ρ = .3	1.153	.029	1.174	.016	1.162	.033
IM, ρ = .6	1.457	.023	1.480	.013	1.480	.035
IM, ρ = .9	1.764	.034	1.804	.019	2.220	.125
PM, ρ = .3	.518	.015	.535	-.001	.538	.031
PM, ρ = .6	.692	.015	.707	.003	.709	.023
PM, ρ = .9	.893	.004	.901	-.001	.900	.034

Presented in Table 9 were the means of item parameters estimated with the CM calibration approach. The composite *a*–parameter was computed to approximate the amount of information a multidimensional item provided. In addition to *a*- and *d*-parameters, the estimated ability correlations were also included in Table 8. For the UNI designs, the CM calibration allotted almost equal discrimination powers to four domains, ranging from .238 to .246, and the composite *a*-parameters (.959 for *N*=5, .981 for *N*=10, and .980 for *N*=20) closely approximated their true values. The CM calibration underestimated *a*-parameters for the IM designs unless the ability correlation was as high as *ρ*=.9 (ranging from .303 to .317 for *N*=5, .366 to .392 for *N*=10, and .430 to .466 for *N*=20), but accurately estimated *a*-parameters in the PM designs regardless of true ability correlations (ranging from .237 to .250 for *N*=5, .238 to .247 for *N*=10, and .238 to .251 for *N*=20). This might be attributable to that fact the complex-structured multidimensional response data as in the IM designs were naturally more difficult to calibrate than the simple-structured response data as in the PM designs. The *d*-parameters were calibrated with acceptably small biases. The correlation coefficients between abilities were generally calibrated with precision for the UNI and PM designs, in which they were almost 1.0 for the UNI designs, and roughly .3, .6, and .9 for PM designs with *ρ*=.3, .6, and .9 respectively. Ability correlations, however, were significantly overestimated for the IM designs.

Table 10 presented the means of item parameters estimated with the SU calibration approach. The discrimination power was evenly allotted into domains in the UNI designs, where *a*-parameters ranged from .239 to .252. The *a*-parameters were underestimated in the IM designs when the correlation between abilities was not high, ranging from .298 to .315 when *ρ*=.3 and .365 to .392 when *ρ*=.6. When abilities were correlated at .9, *a*-parameters were estimated in the vicinity of their true values, ranging from .440 to .471. For the PM designs, *a*-parameters stayed within a narrow range between .238 and .262, regardless of the correlation between abilities. Overall, the measurement information provided by an item, as approximated by the composite *a*-parameter, was overestimated for the IM designs and underestimated for the PM designs. The domain-specific *d*-parameters were generally estimated with precision. In all conditions, differential domain difficulty was observed in estimated item parameters. However, such differences were greater in the UNI design (*d*-parameters ranging from -.156 to .142) and the PM designs (*d*-parameters ranging from -.186 to .170) than in the IM designs (*d*-parameters ranging from -.058 to .075).

Table 9. Estimated Item Parameters in the CM Calibration

	a_1	a_2	a_3	a_4	d	r[1]	a[2]
UNI							
N = 5	.238	.244	.237	.240	-.015	.997	.959
N = 10	.246	.238	.246	.251	-.018	.997	.981
N = 20	.246	.247	.246	.241	.002	.995	.980
IM, ρ = .3							
N = 5	.308	.306	.299	.309	.030	.945	1.222
N = 10	.313	.303	.310	.317	.017	.947	1.242
N = 20	.312	.308	.309	.303	.035	.945	1.233
IM, ρ = .6							
N = 5	.381	.376	.366	.389	.022	.976	1.512
N = 10	.387	.371	.379	.392	.014	.978	1.529
N = 20	.388	.381	.386	.379	.036	.977	1.533
IM, ρ = .9							
N = 5	.450	.450	.430	.452	.034	.995	1.783
N = 10	.461	.441	.455	.466	.019	.995	1.823
N = 20	.466	.453	.459	.450	.164	.995	1.826
PM, ρ = .3							
N = 5	.238	.240	.237	.247	.018	.296	.962
N = 10	.249	.241	.246	.250	-.002	.297	.985
N = 20	.246	.245	.247	.243	.039	.297	.982
PM, ρ = .6							
N = 5	.239	.241	.238	.246	.015	.595	.963
N = 10	.246	.238	.244	.251	.016	.594	.978
N = 20	.245	.247	.247	.244	.027	.594	.983
PM, ρ = .9							
N = 5	.238	.243	.238	.246	.009	.896	.963
N = 10	.246	.238	.245	.251	-.001	.894	.980
N = 20	.245	.245	.247	.241	.035	.895	.978

Note: [1]estimated correlations between abilities, [2]composite *a*-parameters.

Table 10. Estimated Item Parameters in the SU Calibration

	a_1	a_2	a_3	a_4	d_1	d_2	d_3	d_4	a[1]	d[2]
UNI										
N = 5	.24	.24	.24	.24	.14	.05	-.06	-.16	.96	-.02
N = 10	.25	.24	.25	.25	.11	.04	-.05	-.13	.98	-.02
N = 20	.25	.25	.25	.24	.13	.05	-.05	-.13	.98	.00
IM, ρ = .3										
N = 5	.31	.30	.30	.31	.07	.02	-.01	-.05	1.22	.03
N = 10	.31	.30	.31	.32	.05	.02	-.01	-.05	1.23	.02
N = 20	.31	.31	.31	.30	.05	.05	-.02	-.04	1.22	.03
IM, ρ = .6										
N = 5	.38	.38	.37	.39	.07	.02	-.01	-.05	1.52	.02
N = 10	.39	.37	.38	.39	.05	.02	-.01	-.06	1.52	.01
N = 20	.39	.38	.38	.38	.05	.05	-.02	-.04	1.52	.04
IM, ρ = .9										
N = 5	.46	.45	.44	.46	.08	.02	-.01	-.05	1.80	.03
N = 10	.46	.44	.46	.47	.05	.02	-.01	-.06	1.83	.02
N = 20	.47	.45	.46	.45	.05	.05	-.02	-.04	1.83	.04
PM, ρ = .3										
N = 5	.24	.24	.24	.25	.17	.06	-.06	-.16	.96	.02
N = 10	.25	.24	.25	.25	.16	.05	-.06	-.15	.99	.00
N = 20	.25	.25	.25	.24	.13	.06	-.05	-.12	.98	.04
PM, ρ = .6										
N = 5	.24	.24	.24	.25	.17	.06	-.06	-.16	.96	.02
N = 10	.25	.24	.25	.25	.16	.05	-.07	-.15	.99	-.01
N = 20	.25	.25	.25	.25	.13	.06	-.05	-.12	.98	.03
PM, ρ = .9										
N = 5	.24	.24	.24	.26	.17	.06	-.06	-.19	.98	-.01
N = 10	.25	.24	.25	.25	.16	.05	-.06	-.14	.98	.00
N = 20	.25	.25	.25	.24	.14	.06	-.05	-.12	.98	.04

Note: [1]composite *a*-parameter, [2]composite *d*-parameter

Estimated Scores

Table 11 provided a summary of estimated scores using item parameters from the CU calibration. As laid out in the preceding chapter, there were three sets of total score and two sets of subscores computed, including the raw total score, FA composite total score, MSE composite total score, raw subscores, and augmented subscores. The means of both the raw and composite total scores were approximately 11 for the 20-item tests, 22 for the 40-item tests, and 44 for the 80-item tests, and these means remained almost invariant across test designs. Raw subscores remained invariant across test designs too, and in each design they reflected the differential domain difficulties, such as higher mean in Subscore 1 (ARI) and lower mean in Subscore 4 (STA). However, such differences were more evident in the UNI designs and the PM designs than in the IM designs. For example, in the UNI designs, the difference between Subscore 1 and Subscore 4 was 1.2 when *N*=5, 2.0 when *N*=10, and 4.4 when *N*=20, whereas in the IM designs with ρ=.3, the difference was .4, .7, and 1.5 respectively. In other IM designs, this difference was even smaller. The means of augmented subscores were nearly identical to those of raw subscores. Similar patterns regarding the total and subscores estimated with item parameters from the CM and SU calibrations were observed (see Table 12 and Table 13). Thus, they were not discussed in details again.

Table 11. Summary of Estimated Scores Using Item Parameter from the CU Calibration

	Total Score			Raw Subscore				Augmented Subscore			
	Raw	FA	MSE	1	2	3	4	1	2	3	4
UNI											
N = 5	10.8	10.8	10.8	3.3	2.9	2.5	2.1	3.4	3.0	2.6	2.2
N = 10	21.7	21.7	21.7	6.4	5.8	5.1	4.4	6.5	5.9	5.2	4.5
N = 20	43.9	43.9	43.9	13.2	11.8	10.1	8.8	13.2	11.8	10.2	8.9
IM, ρ = .3											
N = 5	10.9	10.9	10.9	2.9	2.8	2.7	2.5	3.0	2.8	2.7	2.6
N = 10	21.9	21.9	21.9	5.8	5.6	5.4	5.1	5.9	5.7	5.4	5.1
N = 20	44.3	44.3	44.3	11.7	11.7	10.6	10.3	11.7	11.7	10.7	10.4
IM, ρ = .6											
N = 5	10.8	10.8	10.8	2.9	2.7	2.7	2.5	3.0	2.8	2.7	2.6
N = 10	21.9	21.9	21.9	5.8	5.6	5.4	5.1	5.8	5.7	5.5	5.2
N = 20	44.1	44.1	44.1	11.6	11.6	10.6	10.4	11.6	11.7	10.7	10.4
IM, ρ = .9											
N = 5	10.8	10.8	10.8	2.9	2.7	2.7	2.6	2.9	2.8	2.7	2.6
N = 10	22.0	22.0	22.0	5.7	5.6	5.4	5.2	5.8	5.7	5.5	5.2
N = 20	43.9	43.9	43.9	11.4	11.5	10.7	10.3	11.4	11.6	10.8	10.3
PM, ρ = .3											
N = 5	11.0	11.0	11.0	3.4	3.0	2.5	2.1	3.5	3.0	2.5	2.1
N = 10	21.9	21.9	21.9	6.8	5.8	5.0	4.2	6.8	5.9	5.0	4.3
N = 20	44.5	44.5	44.5	13.3	12.0	10.2	9.0	13.4	12.1	10.2	9.0
PM, ρ = .6											
N = 5	10.9	10.9	10.9	3.4	3.0	2.5	2.1	3.5	3.0	2.5	2.1
N = 10	21.9	21.9	21.9	6.8	5.9	5.0	4.3	6.8	5.9	5.1	4.3
N = 20	44.3	44.3	44.3	13.2	12.0	10.1	8.9	13.3	12.1	10.2	9.0
PM, ρ = .9											
N = 5	10.9	10.9	10.9	3.4	3.0	2.5	2.1	3.5	3.1	2.5	2.2
N = 10	21.9	21.9	21.9	6.7	5.8	5.0	4.3	6.8	5.9	5.1	4.4
N = 20	44.5	44.5	44.5	13.2	12.0	10.2	9.0	13.3	12.1	10.2	9.0

Table 12. Summary of Estimated Scores with Item Parameters from the CM Calibration

	Total Score			Raw Subscore				Augmented Subscore			
	Raw	FA	MSE	1	2	3	4	1	2	3	4
UNI											
N = 5	10.6	10.6	10.6	3.4	2.9	2.4	2.0	3.4	2.9	2.4	2.0
N = 10	21.4	21.4	21.4	6.5	5.8	5.0	4.2	6.5	5.8	5.0	4.2
N = 20	43.4	43.4	43.4	13.2	11.7	9.9	8.5	13.2	11.7	9.9	8.5
IM, ρ = .3											
N = 5	10.6	10.6	10.6	2.9	2.7	2.6	2.4	2.9	2.7	2.6	2.4
N = 10	21.5	21.5	21.5	5.8	5.6	5.3	4.9	5.8	5.6	5.3	4.9
N = 20	43.7	43.7	43.7	11.6	11.6	10.5	10.1	11.6	11.6	10.5	10.1
IM, ρ = .6											
N = 5	10.5	10.5	10.5	2.9	2.7	2.6	2.4	2.9	2.7	2.6	2.4
N = 10	21.4	21.4	21.4	5.7	5.5	5.3	4.9	5.7	5.5	5.3	4.9
N = 20	43.4	43.4	43.4	11.5	11.5	10.4	10.1	11.5	11.5	10.4	10.1
IM, ρ = .9											
N = 5	10.4	10.4	10.4	2.8	2.6	2.5	2.4	2.8	2.6	2.5	2.4
N = 10	21.5	21.5	21.5	5.7	5.5	5.4	5.0	5.7	5.5	5.4	5.0
N = 20	42.9	42.9	42.9	11.1	11.4	10.4	9.9	11.1	11.4	10.4	9.9
PM, ρ = .3											
N = 5	10.7	10.7	10.7	3.5	3.0	2.4	1.9	3.5	3.0	2.4	1.9
N = 10	21.5	21.5	21.5	6.8	5.8	4.8	4.0	6.8	5.8	4.8	4.0
N = 20	44.1	44.1	44.1	13.3	12.0	10.0	8.7	13.3	12.0	10.0	8.7
PM, ρ = .6											
N = 5	10.7	10.7	10.7	3.5	3.0	2.4	1.9	3.5	3.0	2.4	1.9
N = 10	21.7	21.7	21.7	6.8	5.9	4.9	4.1	6.8	5.9	4.9	4.1
N = 20	43.9	43.9	43.9	13.3	12.0	10.0	8.6	13.3	12.0	10.0	8.6
PM, ρ = .9											
N = 5	10.7	10.7	10.7	3.4	3.0	2.4	1.9	3.4	3.0	2.4	1.9
N = 10	21.5	21.5	21.5	6.8	5.8	4.9	4.0	6.8	5.8	4.9	4.0
N = 20	44.0	44.0	44.0	13.3	12.0	10.0	8.7	13.3	12.0	10.0	8.7

Table 13. Summary of Estimated Score with Item Parameters from the SU Calibration

	Total Score			Raw Subscore				Augmented Subscore			
	Raw	FA	MSE	1	2	3	4	1	2	3	4
UNI											
N = 5	10.5	10.5	10.5	3.3	2.9	2.3	1.9	3.3	2.9	2.3	1.9
N = 10	21.2	21.2	21.2	6.4	5.7	4.9	4.1	6.4	5.7	4.9	4.1
N = 20	43.3	43.3	43.3	13.2	11.7	9.9	8.5	13.2	11.7	9.9	8.5
IM, ρ = .3											
N = 5	10.4	10.4	10.4	2.9	2.7	2.5	2.4	2.9	2.7	2.5	2.4
N = 10	21.3	21.3	21.3	5.7	5.5	5.2	4.9	5.7	5.5	5.2	4.9
N = 20	43.6	43.6	43.6	11.6	11.6	10.4	10.1	11.6	11.6	10.4	10.1
IM, ρ = .6											
N = 5	10.5	10.5	10.5	2.9	2.7	2.6	2.4	2.9	2.7	2.6	2.4
N = 10	21.4	21.4	21.4	5.7	5.5	5.3	4.9	5.7	5.5	5.3	4.9
N = 20	43.4	43.4	43.4	11.5	11.5	10.4	10.1	11.5	11.5	10.4	10.1
IM, ρ = .9											
N = 5	10.4	10.4	10.4	2.8	2.6	2.5	2.4	2.8	2.6	2.5	2.4
N = 10	21.5	21.5	21.5	5.7	5.5	5.4	5.0	5.7	5.5	5.4	5.0
N = 20	42.9	42.9	42.9	11.1	11.4	10.4	9.9	11.1	11.4	10.4	9.9
PM, ρ = .3											
N = 5	10.7	10.7	10.7	3.5	3.0	2.4	1.9	3.5	3.0	2.4	1.9
N = 10	21.5	21.5	21.5	6.8	5.8	4.8	4.0	6.8	5.8	4.8	4.0
N = 20	44.1	44.1	44.1	13.3	12.0	10.0	8.7	13.3	12.0	10.0	8.7
PM, ρ = .6											
N = 5	10.7	10.7	10.7	3.5	3.0	2.4	1.9	3.5	3.0	2.4	1.9
N = 10	21.7	21.7	21.7	6.8	5.9	4.9	4.1	6.8	5.9	4.9	4.1
N = 20	43.9	43.9	43.9	13.3	12.0	10.0	8.6	13.3	12.0	10.0	8.6
PM, ρ = .9											
N = 5	10.7	10.7	10.7	3.4	3.0	2.4	1.9	3.4	3.0	2.4	1.9
N = 10	21.5	21.5	21.5	6.8	5.8	4.9	4.0	6.8	5.8	4.9	4.0
N = 20	44.0	44.0	44.0	13.3	12.0	10.0	8.7	13.3	12.0	10.0	8.7

Figure 2 and 3 displayed score distributions estimated with item parameters from the CU calibration in one simulation replication. In Figure 2, raw total scores were estimated with almost equal medians in all designs but larger variance in the IM designs and smaller variance in the PM designs. Due to the differential domain difficulties, the subscore medians differed across domains in each design, yet the order agreed with that of hypothetical domain difficulties. The variances of raw subscores were smaller in the PM designs and larger in the IM designs. Some subscore outliers appeared in the PM designs, especially when ρ=.3. In Figure 3, the MSE composite total score failed in the CU calibration, showing eerie distributions with unrealistic outliers (extremely large value or negative value) throughout all designs. This was possibly because of multicollinearity presented among correlated subscores, to which Kelley's true score regression that Longford's score-aggregation method was based on was vulnerable. The same problem was observed in augmented subscores in the UNI and PM designs, in which the score distributions went awry possibly because of multicollinearity.

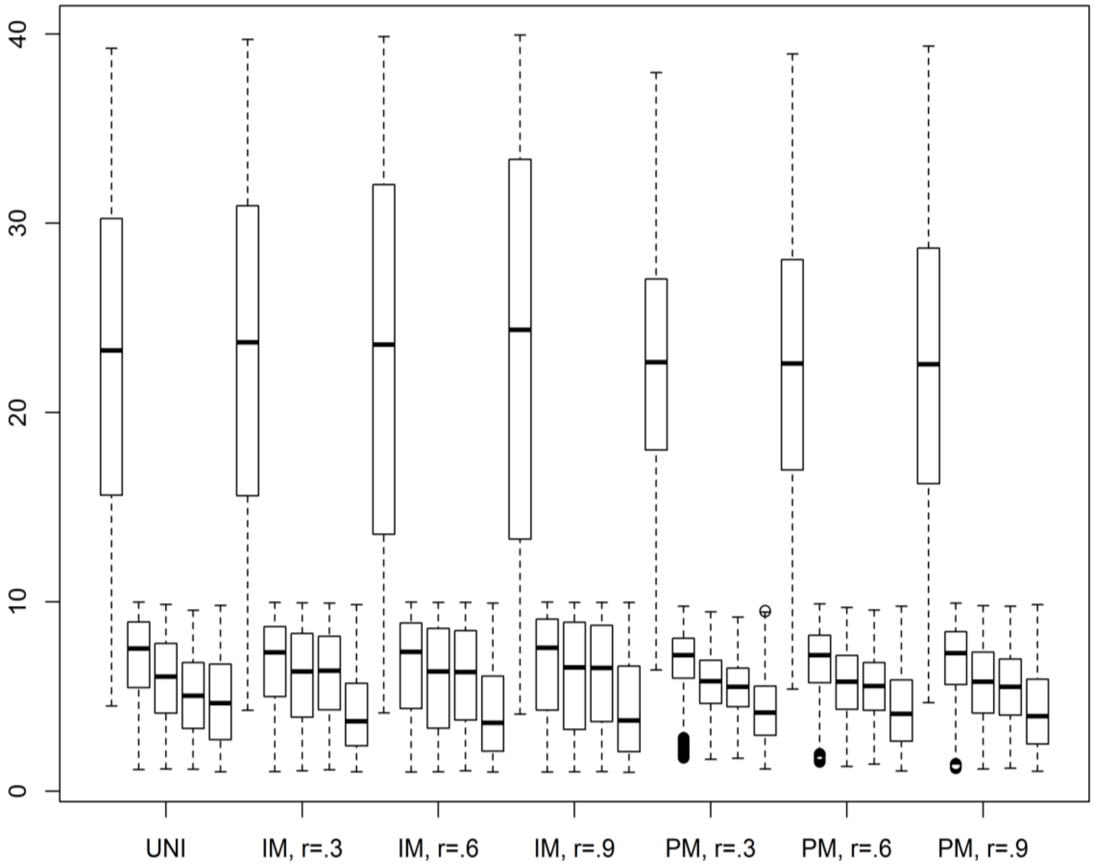

Figure 2. Distributions of Raw Scores for the CU Calibration when *N*=10

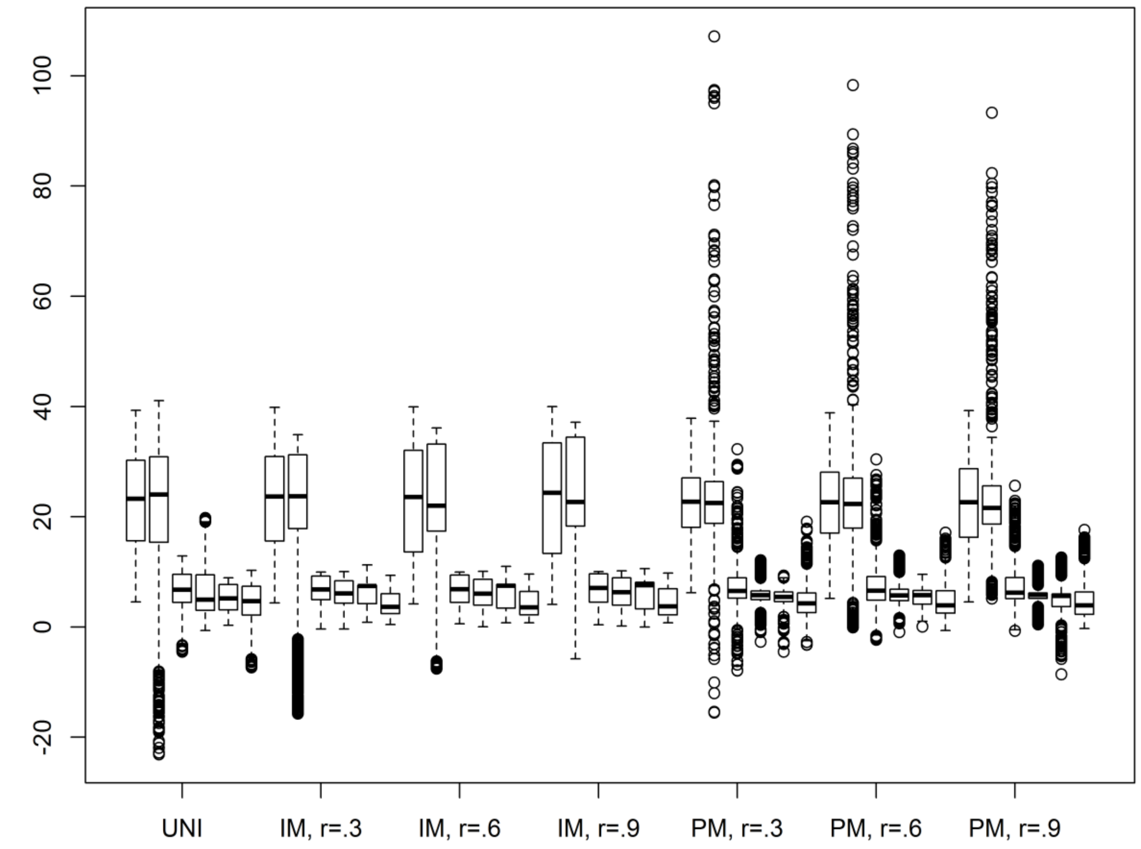

Figure 3. Distributions of Alternative Scores for the CU Calibration when *N*=10

Figure 4 and 5 displayed score distributions estimated with item parameters from the CM calibration from one simulation replication. In Figure 4, raw total scores were estimated with slightly higher medians and smaller variance in the PM designs than in the UNI and IM designs. The score variance also increased as the true ability correlations increased in the IM and PM designs. For raw subscores, distributions were rather leptokurtic in the IM designs. The medians of subscores varied across domains, yet the order did not always agreed with the hypothetical domain difficulties (e.g. in the IM designs). In Figure 5, however, alternative scores showed more reasonable distributions in the CM calibration, compared with the CU calibration. Both alternative total scores and subscores showed similar distributions with scores in Figure 4. Figure 6 and 7 showed similar pattern of score distributions estimated with item parameters from the SU approach. Thus, they were not discussed in details again.

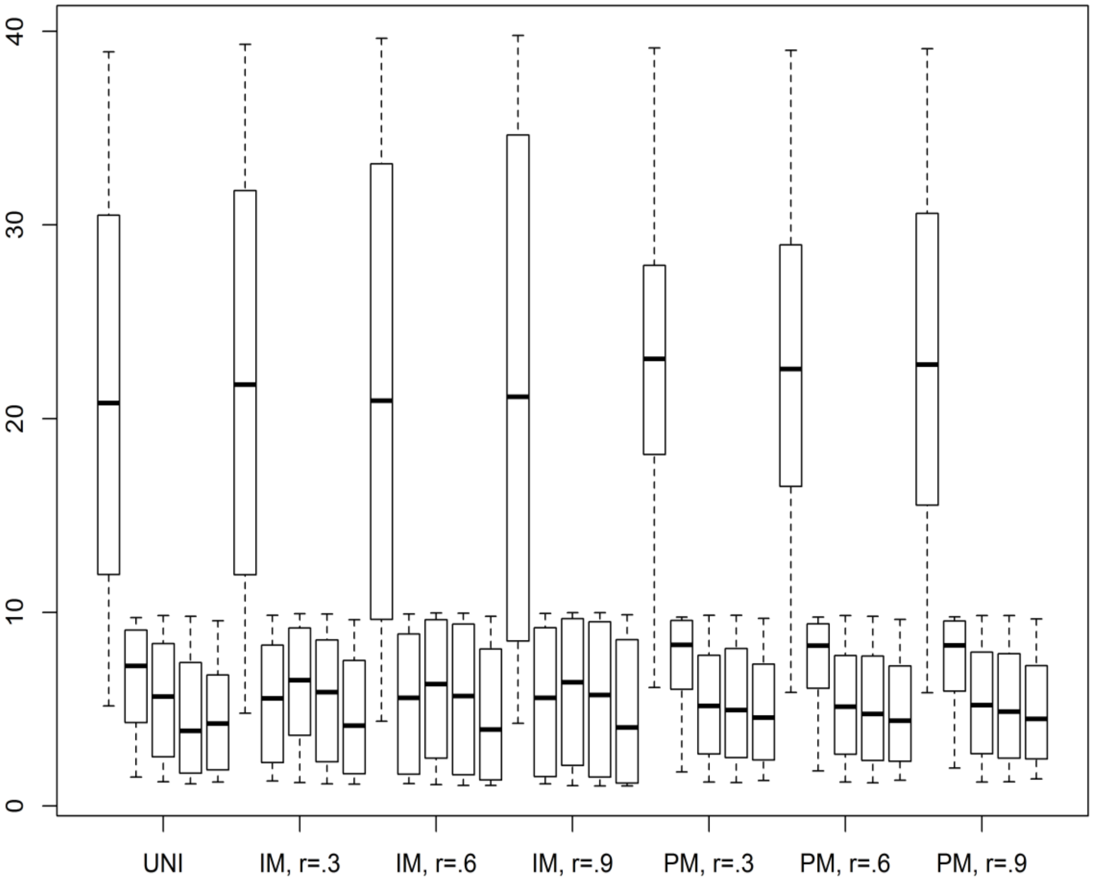

Figure 4. Distributions of Raw Scores for the CM Calibration when *N*=10

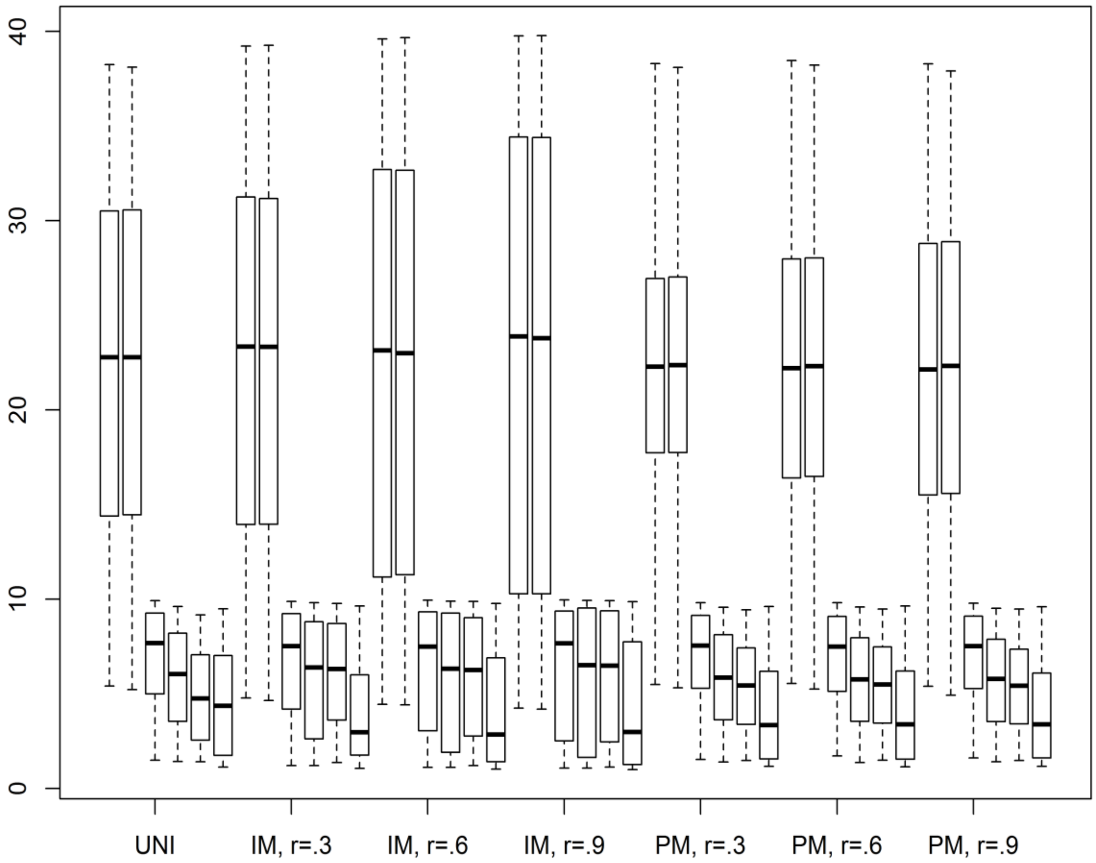

Figure 5. Distributions of Alternative Scores for the CM Calibration when *N*=10

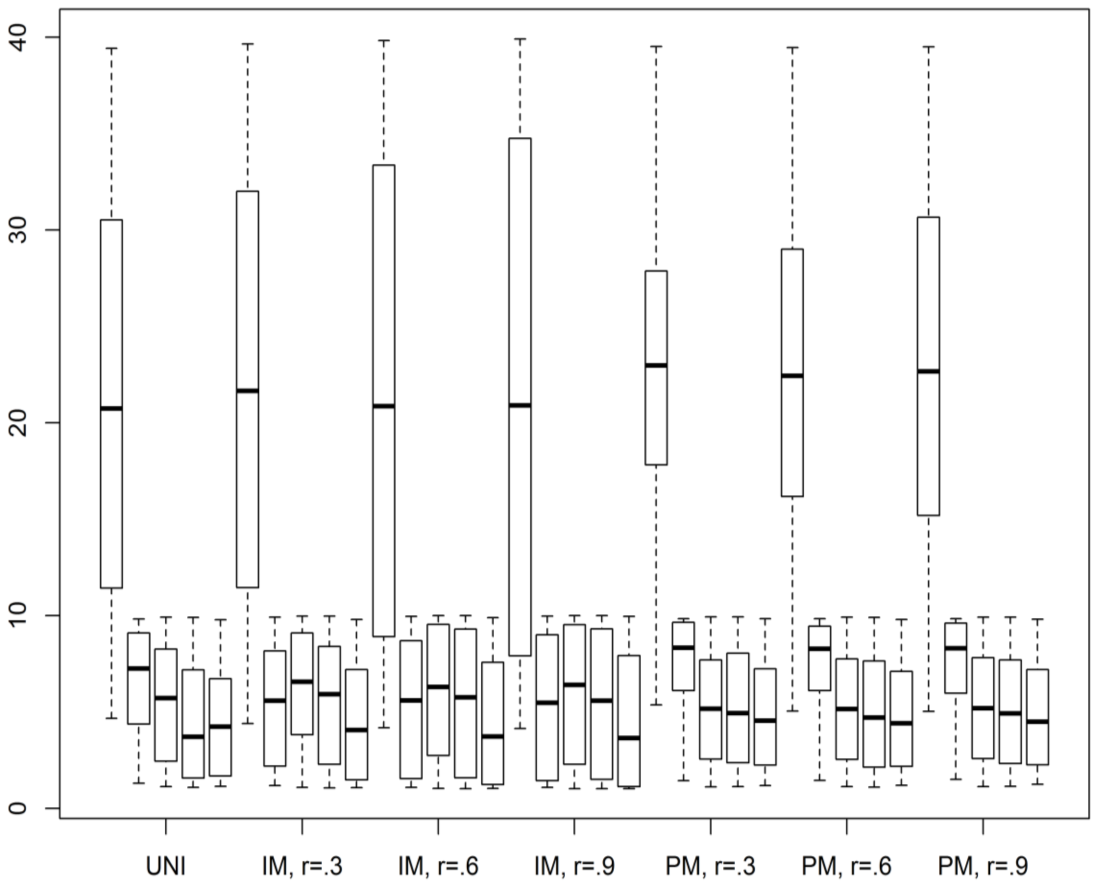

Figure 6. Distributions of Raw Scores for the SU Calibration when *N*=10

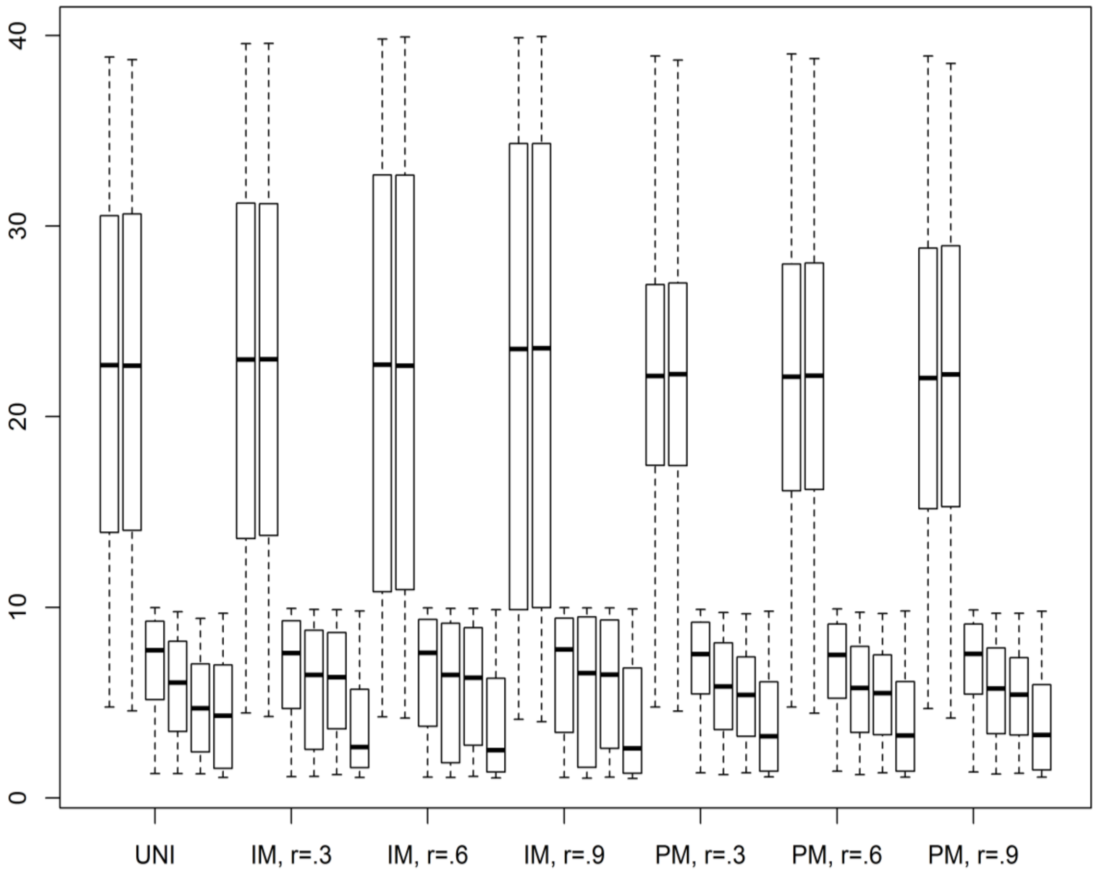

Figure 7. Distributions of Alternative Scores for the SU Calibration when *N*=10

Table 14 provided a summary of correlations between estimated subscores, averaged over 30 replications in each condition. As they should, subscores estimated with the CU calibration were highly correlated with one another, at least .97. Subscores were correlated at a much lower degree for the CM (ranging from .56 to .90 for the UNI and IM designs, .16 to .75 for the PM designs) and SU calibrations (ranging from .56 to .92 for the UNI and IM designs, .16 to .75 for the PM designs). It was worth noting that subscores generally correlated at a lower degree in the PM designs than in the IM designs, possibly because of the higher degree of multidimensionality in the PM designs.

Table 14. Correlations between Estimated Subscores

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.965	.977	.977	.561	.731	.832	.555	.728	.835
IM, ρ=.3	.981	.989	.993	.620	.757	.829	.614	.755	.839
IM, ρ=.6	.980	.988	.993	.691	.815	.871	.684	.818	.891
IM, ρ=.9	.981	.988	.993	.743	.847	.893	.738	.854	.921
PM, ρ=.3	.975	.979	.987	.162	.211	.244	.161	.210	.245
PM, ρ=.6	.967	.975	.983	.327	.422	.491	.323	.421	.493
PM, ρ=.9	.962	.969	.980	.498	.643	.747	.492	.640	.750

Table 15 presented the average disattenuated correlations between estimated subscores of four domains. The disattenuated correlations between two subscores indicated the degree of associations between two subscores when the measurement errors were eliminated. It was computed as follows:

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} \tag{4.1}$$

where ρ_{xy} is the correlation between variable X and variable Y, $\rho_{xx'}$ is the reliability of variable X, and $\rho_{yy'}$ is the reliability of variable Y. Subscores given by the CU calibration showed disattenuated correlations greater than 1.0 in all conditions, suggesting too much error was contained in estimated subscores. As the test length increased, the disattenuated correlations became less extreme, moving toward 1.0, because the longer test contained less error. Estimated subscores given by both the CM and SU calibrations implied an effect of test design. For the test with UNI design, subscores were correlated at roughly 1.0, indicating the high similarity between subscores. This was an expected result for the unidimensional test. For the test with IM design, subscores were also highly correlated at almost 1.0, suggesting that the unprincipled multidimensionality could not make subscores distinct from each other sufficiently. Conversely, for the test with PM design, disattenuated correlations reflected the true distinctiveness of subscores by showing disattenuated correlations being about .3, .6, and .9 when ρ's were .3, .6, and .9 respectively. Comparing the IM and PM designs, it was evident without a principled design, the multidimensional test failed to present the distinctiveness of subscores.

Table 15. Disattenuated Correlations between Estimated Subscores

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	1.738	1.347	1.160	1.005	1.007	.988	.994	1.002	.992
IM, ρ=.3	1.507	1.249	1.123	.951	.956	.937	.941	.953	.949
IM, ρ=.6	1.388	1.189	1.093	.978	.980	.959	.970	.983	.981
IM, ρ=.9	1.319	1.158	1.078	.998	.991	.970	.995	1.000	.999
PM, ρ=.3	1.763	1.352	1.172	.293	.291	.290	.290	.289	.292
PM, ρ=.6	1.751	1.346	1.167	.590	.583	.583	.583	.581	.585
PM, ρ=.9	1.737	1.340	1.163	.897	.888	.887	.886	.884	.890

Score Accuracy

Figure 8 to 10 showed the correlations and RMSEs between true and estimated raw total scores for the CU, CM, and SU calibration approaches respectively. It should be remembered that the scale of RMSE was five times shorter in graphs than the original scale in order to align with the scale of correlation. For example, if the RMSE read .4 in the graph, the real value of this RMSE was 2.0. According to the graphs, the total score was accurately estimated in all conditions and with all calibration approaches. The estimated total scores were correlated with true total scores at at least .85, .90, and .95 for the short tests (*N*=5), medium tests (*N*=10), and long tests (*N*=20), and the RMSEs were at most 2.1, 3.0, and 4.7 for the short, medium and long tests in all conditions. Considering the score scale, these RMSEs were acceptably small.

Also, as expected, longer tests produced more accurate scores. Correlations became consistently higher, as the test length increased. The evaluation of the RMSEs demanded some adjustments, due to the inequality of score scales. In longer test, the score scale was stretched to have wider range and higher variance. Therefore, the RMSE of the long test (*N*=20) appeared significantly greater than the short and medium test. A simple adjustment was to divide the RMSEs of the long tests (*N*=20) and medium tests (*N*=10) by 4 and 2 so that they were converted onto the scale of the short tests (*N*=5). After adjustments, the least RMSEs were observed for the long tests, followed by the medium tests and short tests. For example, raw RMSEs were, on average, approximately 1.9, 2.7, and 3.9 when *N*=5, *N*=10, and *N*=2, yet adjusted RMSEs were roughly 1.9, 1.3, and 1.0 respectively.

Another significant finding following these graphs was the equality of different calibration approaches. All of three calibration approaches functioned rather comparably to one another in terms of estimating total scores. The theoretical advantages of multidimensional calibrations were not present in these results.

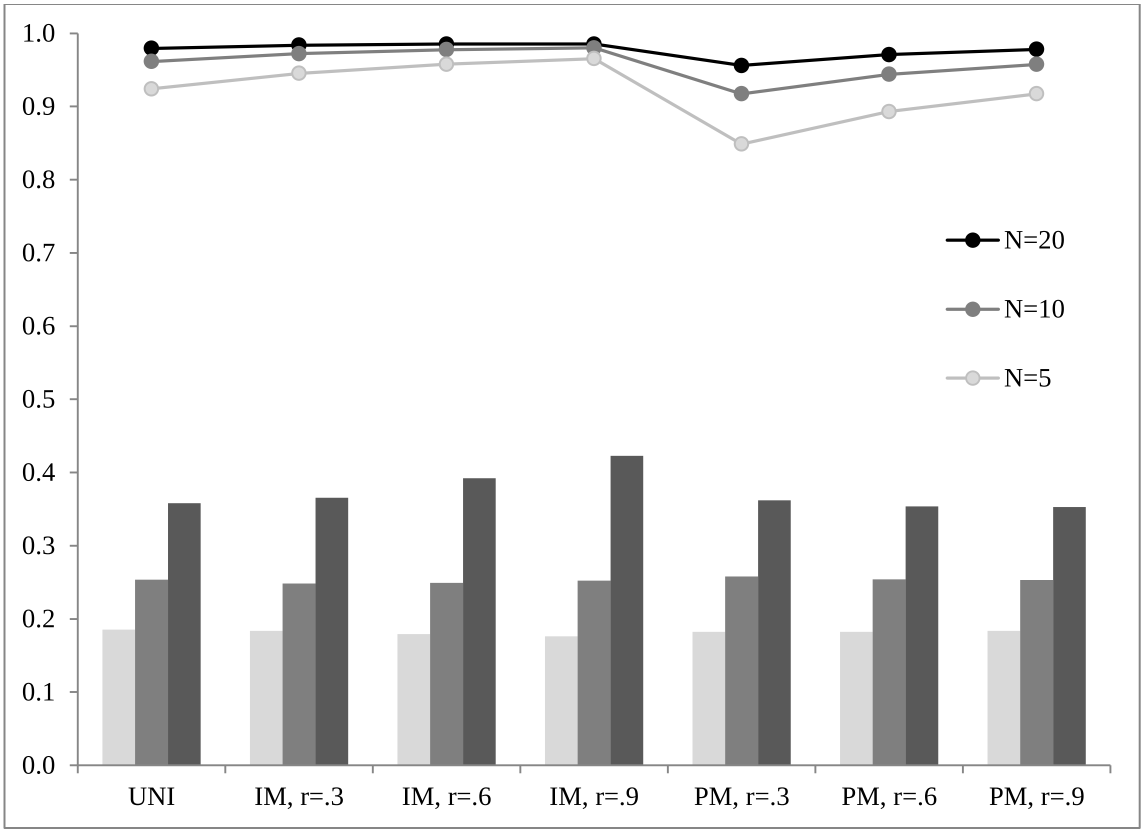

Figure 8. Correlations and RMSEs between True and Estimated Raw Total Score for the CU Calibration

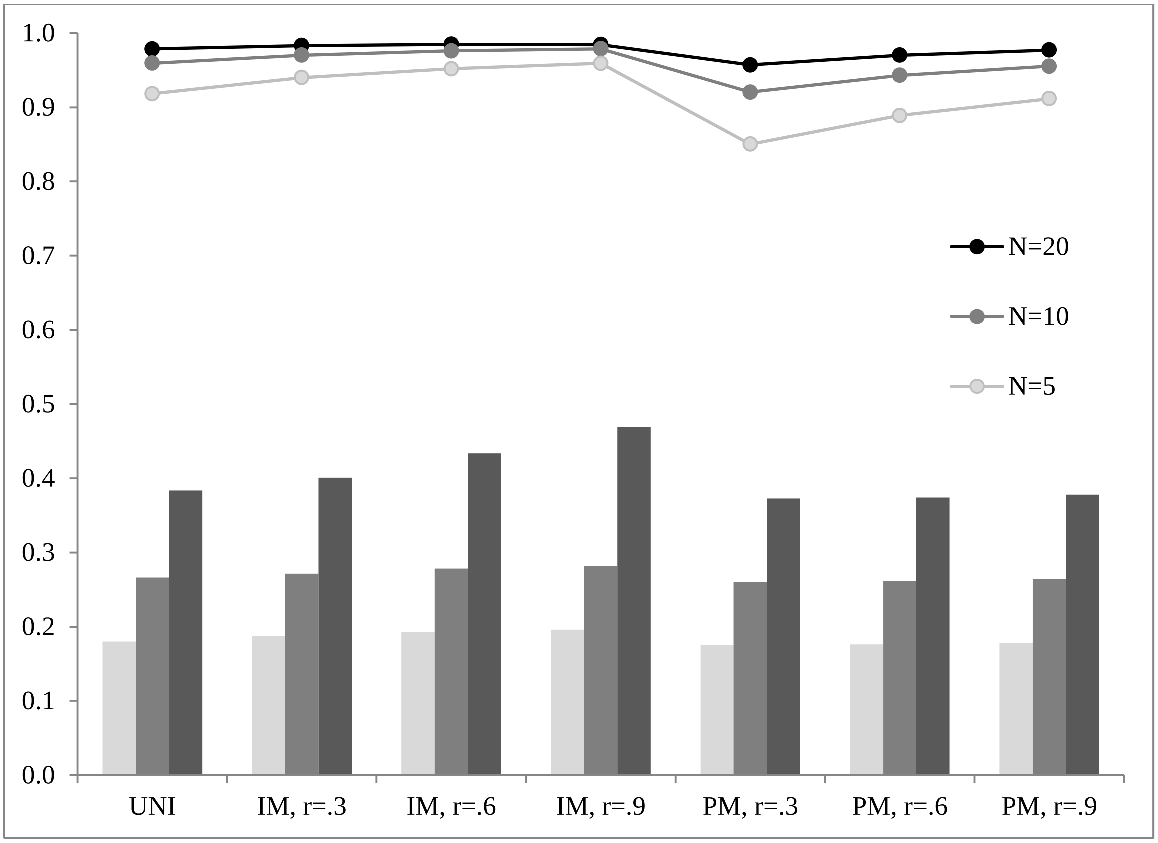

Figure 9. Correlations and RMSEs between True and Estimated Raw Total Scores for the CM Calibration

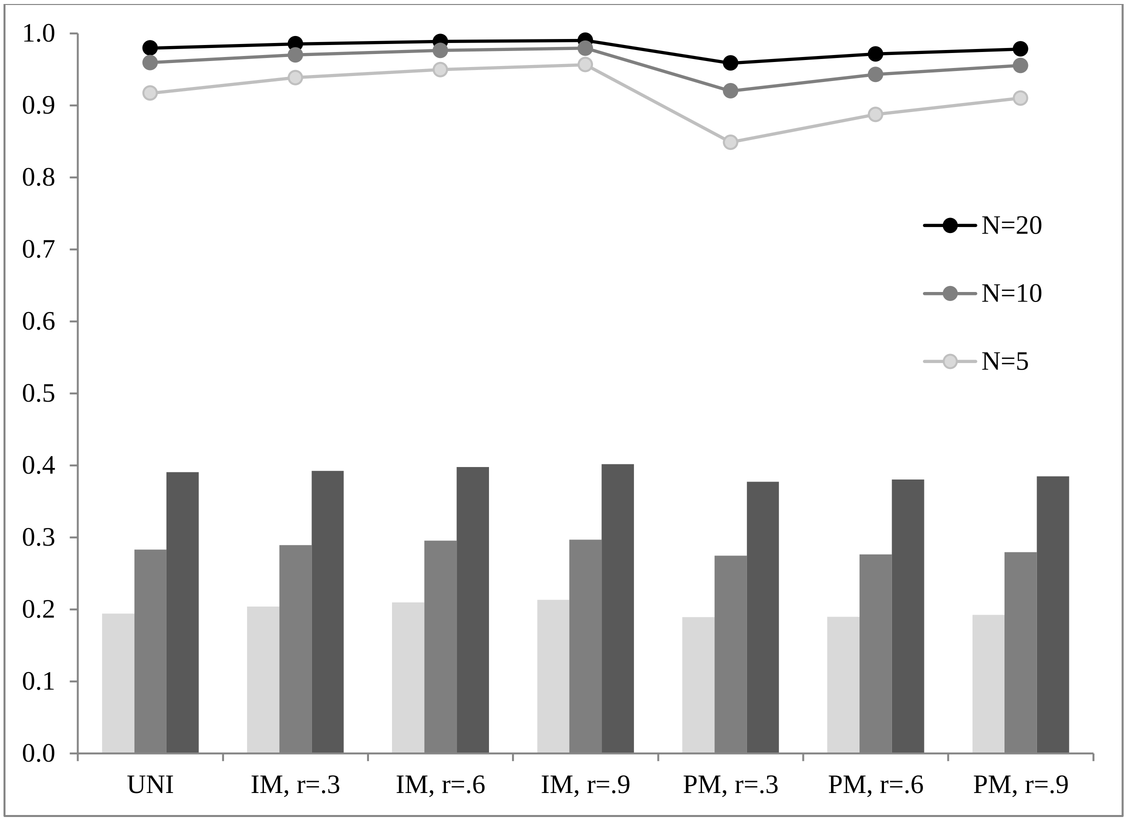

Figure 10. Correlations and RMSEs between True and Estimated Raw Total Scores for the SU Calibration

Figure 11 to 13 presented correlations and RMSEs between true and estimated raw subscores for the CU, CM, and SU calibrations. For brevity, only results regarding Subscore 1 were presented, because of the high similarity among four subscores. With the CU calibration (see Figure 11), subscores were accurately recovered with high correlations (at least .90) and low RMSEs (roughly .52 when *N*=5, .74 when *N*=10, and 1.19 when *N*=20) for the UNI and IM designs. For the PM designs, the quality of score estimation varied as a function of the multidimensionality of the test. When the test was highly multidimensional (*ρ*=.3) and short (*N*=5), the estimated subscores were correlated with true subscores at only .58 with a RMSE of .8. Conversely, when the test was almost unidimensional (ρ=.9) and long (N=20), the estimated subscores were correlated with true subscores at .94 with a RMSE of 1.4 (equal to .35 on the score scale of N=5).

With the CM and SU calibrations (see Figure 12 and 13), the effect of test length became even stronger, especially for the UNI and PM designs. For instance, the correlation between true and estimated subscores differed by roughly .15 between the long and short tests for the UNI and PM designs in the CM and SU calibrations, compared to about .6 in the CU calibration. The effect of test length also interacted with the effect of test design. In the IM designs, for example, the difference between the long and short tests declined from roughly .12 to .08, as the multidimensionality decreased from ρ =.9 to ρ=.3.

Comparing across test designs, subscores were best estimated in the IM designs, with the correlations ranging from .83 to .88 when N=5, .90 to .93 when N=10, .94 to .95 when N=20 and the RMSEs being approximately .89 when N=5, 1.3 when N=10, and 2.0 when N=20. When the test was short (N=5), subscores were generally more poorly estimated with the CM and SU calibrations than the CU calibration. When the test was long enough, the difference between calibrations became practically negligible.

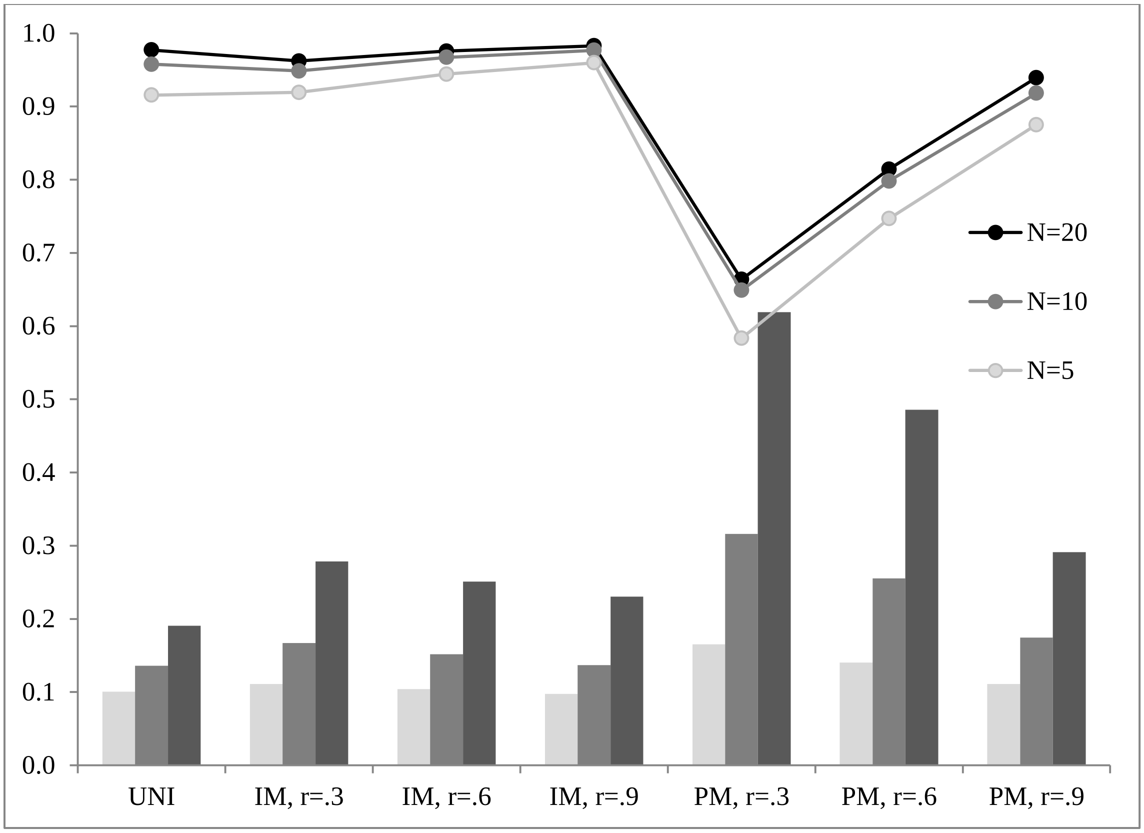

Figure 11. Correlations and RMSEs between True and Estimated Raw Subscore 1 for the CU Calibration

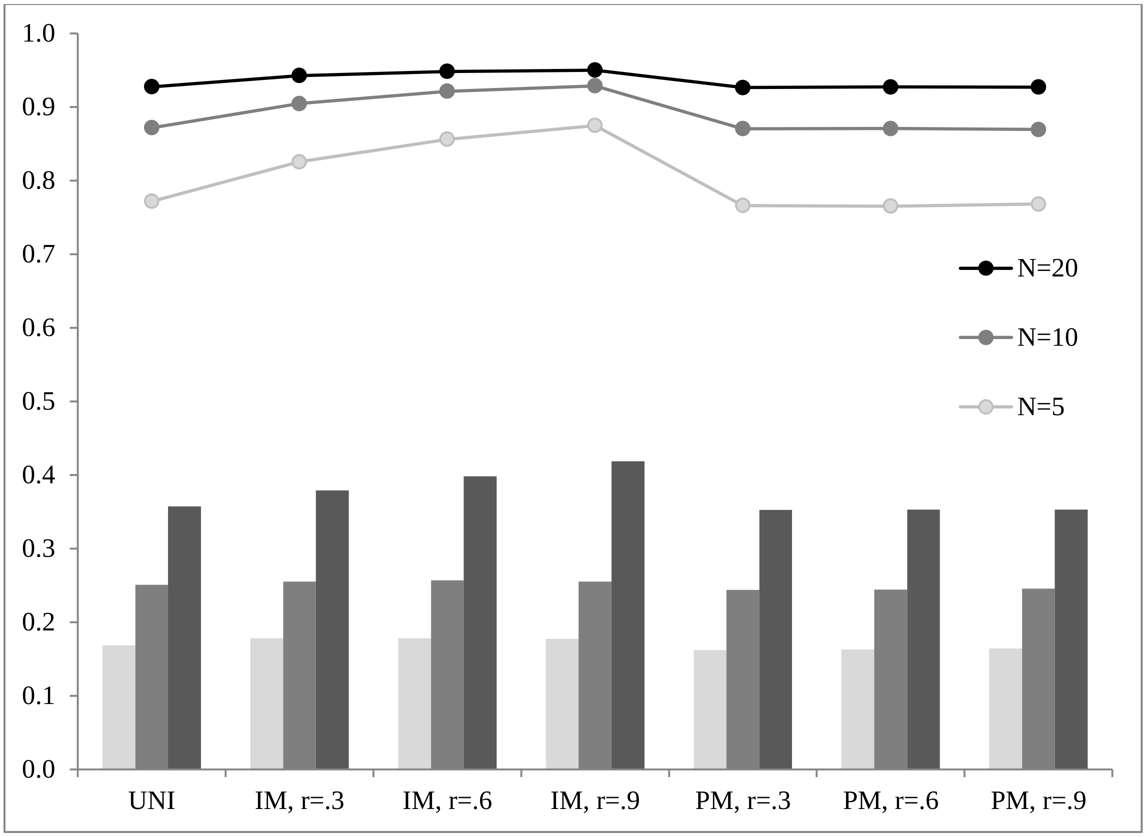

Figure 12. Correlations and RMSEs between True and Estimated Raw Subscore 1 for the CM calibration

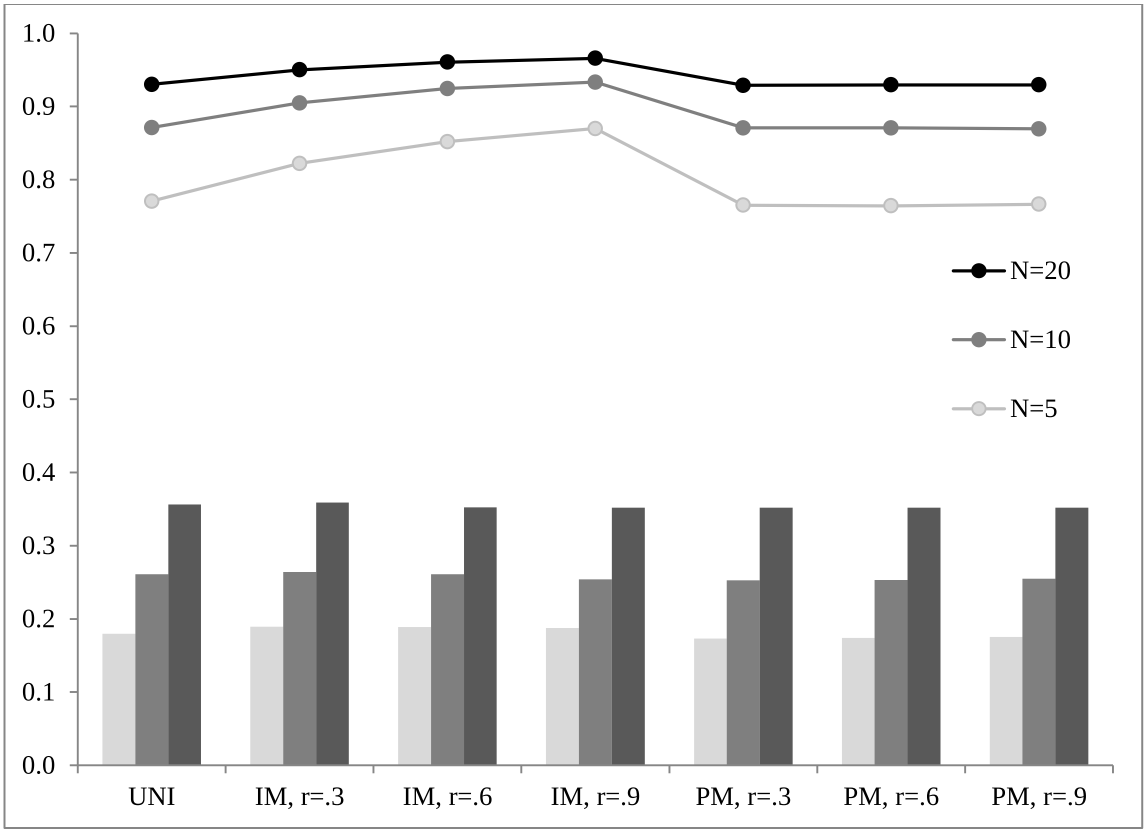

Figure 13. Correlations and RMSEs between True and Estimated Raw Subscore 1 for the SU Calibration

While preceding graphs compared raw total score and subscores individually, the following graphs provided visualized comparisons of all scores within each design in an attempt to investigate the performances of scoring methods and calibration approaches within a certain design. Only correlations were shown below, given that RMSEs would present practically identical information. For the UNI design (see Figure 14), both raw total score and subscores were accurately estimated with the CU calibration, nearly perfectly correlated with true scores. Raw scores were also satisfactorily calibrated with the CM and SU calibration when test was long enough (*N*=20), in which total scores were correlated with true total scores nearly perfectly and subscores were correlated with true subscores at more than .95; however, when test was short (*N*=5), the correlations between true and estimated scores fell significantly down to less than .80 in the CM and SU calibrations, compared to almost 1.0 in the CU calibration. This might be because that the CM calibration overparameterized the response data and the SU calibration utilized insufficient data. In general, alternative scoring methods either malfunctioned (e.g., MSE composite total score and augmented subscores with the CU calibration) or barely showed advantages over raw scores.

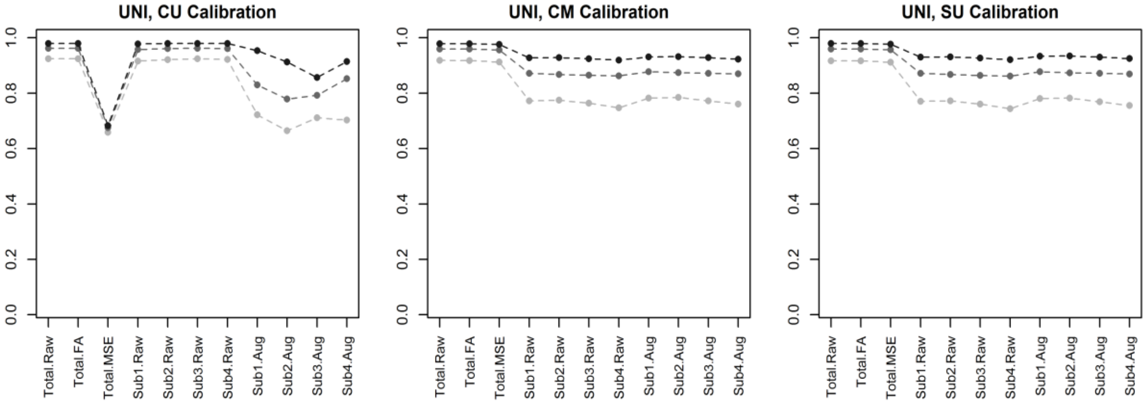

Figure 14. Correlations between True and Estimated Scores for the UNI Design

Figure 15 to 17 presented score comparisons for the IM designs with *ρ*=.3, .6 and .9 respectively. Both raw total score and subscores were accurately estimated with the CU calibration in all of three IM designs. The performance of alternative scoring methods was improved as the multidimensionality of the test decreased from *ρ*=.3 to *ρ*=.9, but they still hardly presented any advantages over traditional scoring methods. A similar effect of the multidimensionality was observed for the CM and SU calibrations. That is, estimated score showed higher correlation with true scores, especially in the short test, as dimensions of the test converged toward the unidimension. Raw scores estimated with the CM and SU calibrations were still less accurate than scores estimated with the CU calibration. Again, alternative scores failed to show significant improvement over raw scores.

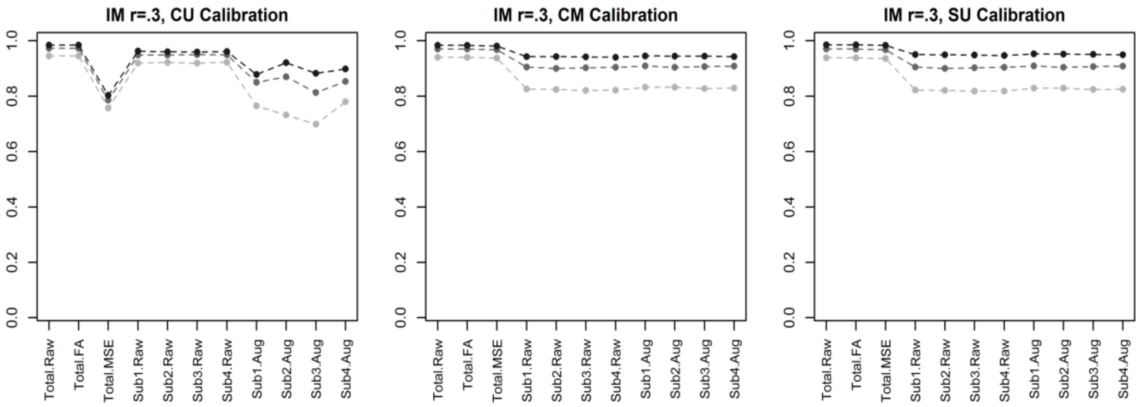

Figure 15. Correlations between True and Estimated Scores for the IM Design with ρ=.3

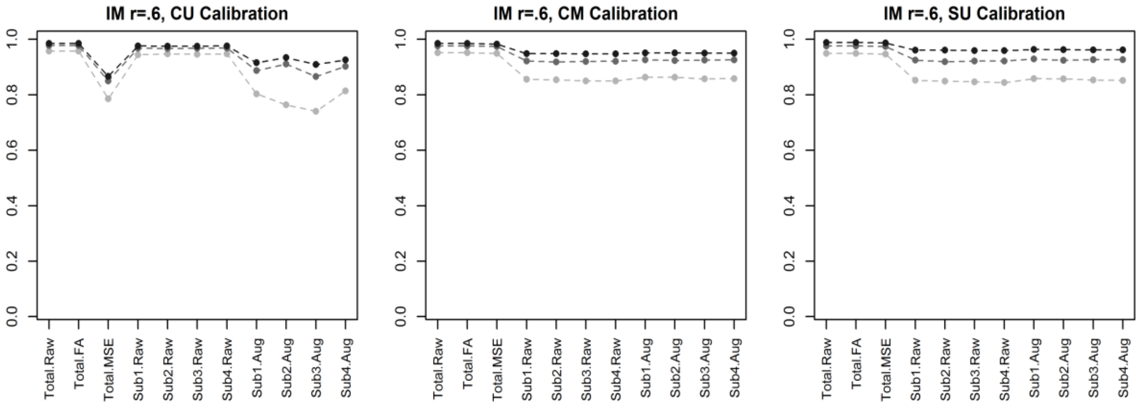

Figure 16. Correlations between True and Estimated Scores for the IM Design with ρ=.6

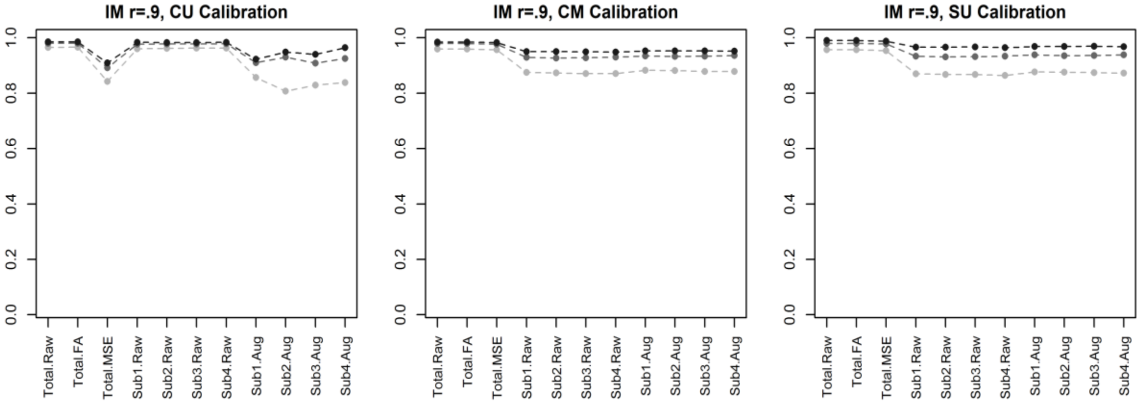

Figure 17. Correlations between True and Estimated Scores for the IM Design with *ρ*=.9

Figure 18 to 20 compared scores for the PM designs with *ρ*=.3, .6 and .9. When the test was designed with multiple distinct dimensions (*ρ*=.3), subscores were poorly estimated with the CU calibration (correlating with true subscores at roughly .6), but nicely estimated with the CM calibration and SU calibration (correlating with true subscores at roughly from.70 to .90), which suggested that the multidimensionality could be recovered with the multidimensional or separate unidimensional calibration approaches but not the conventional unidimensional calibration approach. As it should, the score estimation was significantly improved as the multidimensionality of the test diminished from *ρ*=.3 to *ρ*=.9. For instance, in Figure 20 (the PM design with *ρ*=.9), the results looked almost identical to those in the UNI designs. Again, alternative scores either malfunctioned or barely presented advantages over raw scores.

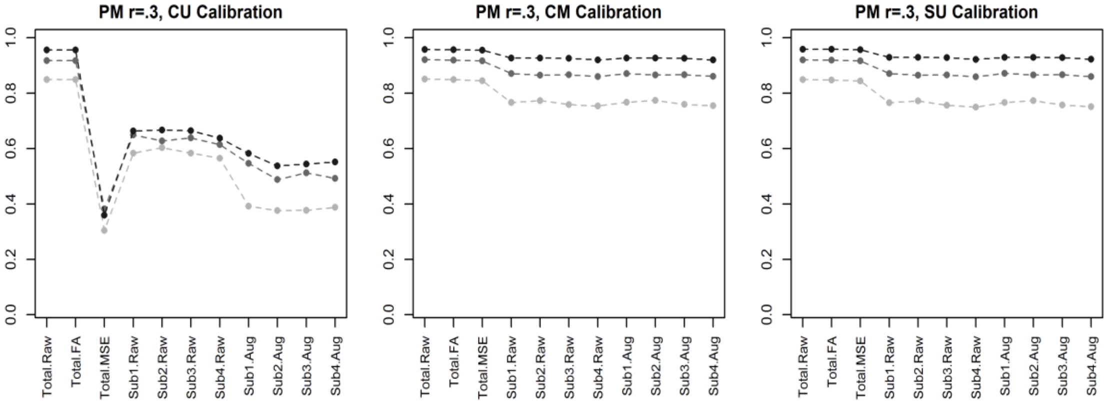

Figure 18. Correlations between True and Estimated Scores for the PM Design with *ρ*=.3

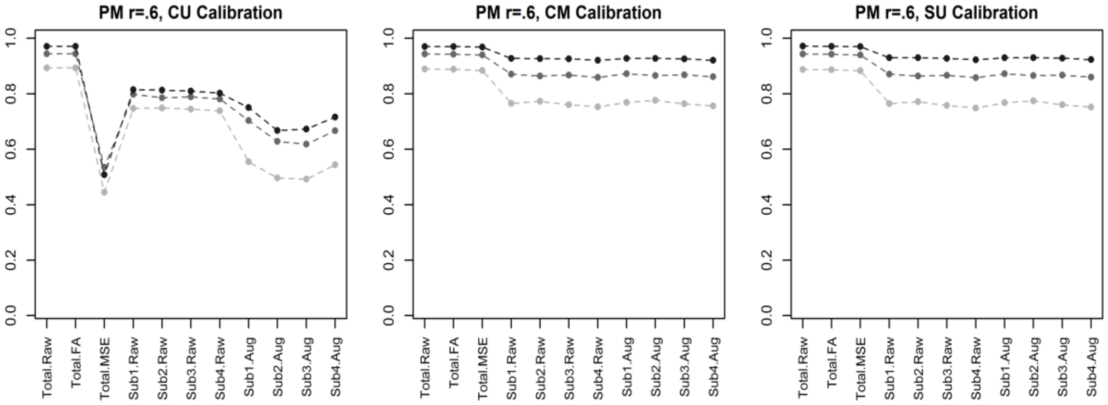

Figure 19. Correlations between True and Estimated Scores for the PM Design with *ρ*=.6

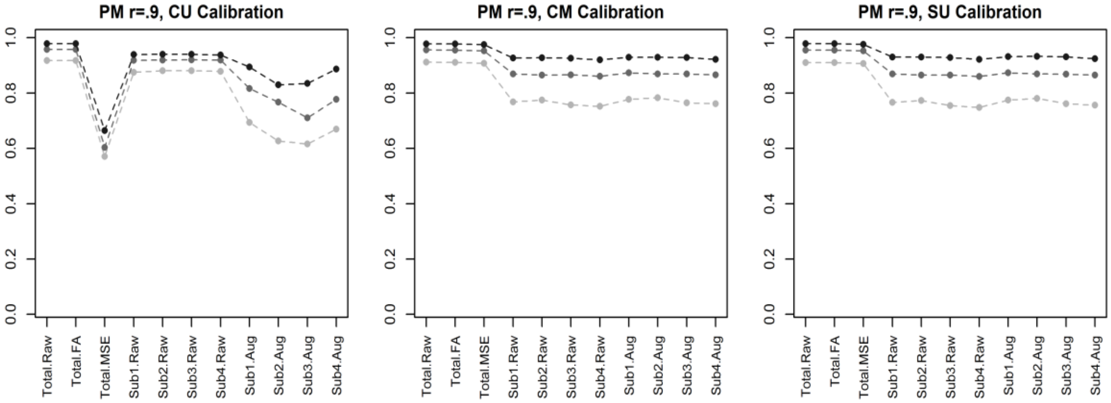

Figure 20. Correlations between True and Estimated Scores for the PM Design with *ρ*=.9

Score Reliability

Table 16 and 17 presented reliability coefficients for raw total score and subscores. As described in Chapter III, since true scores were known from the data generation, the reliability coefficient was computed as the ratio of the true score variance to the observed score variance. In Table 16, as it should, the reliability of the total score was a function of test length, that is, higher reliability observed for longer tests. Comparing across test designs, the total score in the UNI design (at least .84) and IM designs (at least .88) showed satisfactory reliabilities, yet unsatisfactory reliabilities in the PM designs (as low as .72). In addition, the effect of test dimensionality in the multidimensionality designs also interacted with the effect of test length. When dimensions were loosely correlated (*ρ*=.3), the difference between the short test and long test was approximately .9 for the IM designs and 1.9 for the PM designs. When dimensions were tightly correlated (*ρ*=.9), the difference was roughly .5 for the IM designs and 1.2 for the PM designs.

Since subscores were rather homogeneous, only subscore 1 (ARI) was shown in Table 17 as an example. The reliability of subscores was sensitive to test length and test design too. That is, higher reliability was observed for long tests, as predicted by the Spearman-Brown prophecy. Comparing across test designs, the reliabilities were lower in the PM designs than in the UNI and IM designs. Subscore reliability was affected by the calibration approach. In general, the CM and SU calibrations produced appreciably lower reliability than the CU calibrations. When the test was long enough, however, the reliability in the CM and SU calibrations rose to an acceptable level.

Table 16. Reliability of Total Score

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.854	.925	.958	.843	.920	.958	.841	.920	.958
IM, r=.3	.893	.945	.968	.884	.941	.966	.880	.941	.970
IM, r=.6	.918	.956	.970	.906	.953	.970	.901	.953	.978
IM, r=.9	.931	.960	.970	.920	.958	.968	.914	.958	.980
PM, r=.3	.721	.841	.914	.723	.846	.916	.721	.846	.920
PM, r=.6	.797	.891	.943	.790	.889	.941	.787	.889	.943
PM, r=.9	.841	.916	.956	.832	.912	.955	.828	.912	.956

Table 17. Reliability of Subscore 1 (ARI)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.838	.917	.955	.596	.760	.860	.594	.759	.865
IM, r=.3	.845	.900	.926	.681	.818	.888	.676	.819	.903
IM, r=.6	.891	.935	.952	.733	.849	.899	.726	.855	.923
IM, r=.9	.921	.954	.966	.765	.862	.902	.757	.871	.933
PM, r=.3	.340	.421	.441	.587	.758	.858	.585	.758	.863
PM, r=.6	.558	.637	.663	.586	.758	.860	.584	.758	.864
PM, r=.9	.766	.843	.882	.590	.756	.859	.588	.756	.864

Subscore Utility

In addition to accuracy, subscores should also be evaluated by their utilities. Table 18 presented the chances that raw subscores showed added value to raw total score, by the criterion of PRMSE, in 30 replications of each condition. Because all four subscores showed added values in all conditions, only results regarding the CU calibration was showed. It was interesting to observe that the utility of subscore was a function of calibration approach instead of test design. In general, subscores estimated with the CU calibration were unlikely to supplement the total score with added diagnostic information. When the test was short (*N*=5), Subscore 1 and 4 would be worth reporting in at most 28% and 17% of cases, while the other two subscores barely provided added diagnostic information to the total score. When the test was extended to 10 and 20 items per domain, Subscore 1 and 4 were more likely to provide added value in the UNI design and PM designs (at most 59%), whereas the other two subscores still struggled to be diagnostically informative (at most 11%). This might be because that Subscore 1 and 4 were most likely heterogeneous to the total score, as they were, on average, the easiest and hardest domains. In the IM designs, subscores were diagnostically informative in at most 14.3% of cases.

Table 18. Percentage of Subscores Showing Add-valued in the CU Calibration

	N=5, %				N=10, %				N=20, %			
	1	2	3	4	1	2	3	4	1	2	3	4
UNI	20.0	.0	.0	16.7	30.0	.0	.0	20.0	58.6	.0	6.9	48.3
IM, ρ=.3	6.7	6.7	3.3	10.0	3.3	.0	10.0	6.7	10.0	3.3	.0	6.7
IM, ρ=.6	10.7	7.1	7.1	10.7	3.6	.0	7.1	7.1	14.3	3.6	.0	7.1
IM, ρ=.9	6.9	6.9	3.4	6.9	.0	.0	10.0	10.0	14.3	.0	.0	14.3
PM, ρ=.3	20.0	.0	.0	10.0	34.5	.0	.0	24.1	21.4	10.7	3.6	42.9
PM, ρ=.6	23.3	.0	.0	10.0	41.4	.0	.0	24.1	46.7	6.7	3.3	53.3
PM, ρ=.9	27.6	.0	.0	10.3	44.8	3.4	3.4	27.6	44.4	7.4	3.7	51.9

Calibration Model Comparison

Presented in Table 19 were percentages of Q3 outliers in all conditions. For the UNI designs, the percentages of Q3 outliers were at 25.4%, 10.2% and 9.1% for the CU, MC and SU calibration, suggesting that the latent abilities that drove test performance were adequately modeled by the calibration models. For the IM designs, the CU calibration showed a relatively large percentage of Q3 outliers, meaning that the response data were insufficiently modeled. In the best case where ρ=.3 and N=5, there were 26.6% Q3 outliers, while in the worst case where ρ=.9 and N=20, there were 53.7% outliers. Conversely, both the CM and SU calibration resulted in less than 7% outliers. For the PM designs, the response data were adequately modeled with the CM and SU calibration, with less than 10% outliers, whereas the percentage of outliers in the CU calibration could go up to 21%. Overall, as expected, the CM calibration left minimal amount of dependence in items, because it fit the data with more parameters. Interesting, the SU calibration achieved equivalent performance with the CM calibration but used less parameters, probably because the individual modeling of each subtest gave more flexibility to the model than the concurrent unidimensional modeling. However, it should be remembered that better model fit was not the ultimate goal, and an overfitting model could even lead to the restriction on the generalizability.

Table 19. Percentages of Q3 Outliers

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.086	.171	.254	.102	.034	.006	.091	.030	.007
IM, ρ=.3	.151	.241	.311	.081	.023	.003	.065	.018	.006
IM, ρ=.6	.266	.363	.452	.056	.010	.001	.037	.009	.004
IM, ρ=.9	.367	.451	.537	.035	.005	.001	.023	.006	.002
PM, ρ=.3	.011	.084	.216	.103	.034	.006	.092	.029	.007
PM, ρ=.6	.019	.055	.143	.103	.034	.006	.092	.029	.007
PM, ρ=.9	.066	.136	.208	.103	.032	.006	.094	.028	.007

Table 20 and 21 presented the AIC and BIC for calibration approaches in all conditions. According to AIC (see Table 19), the CM calibration fit the response data best for the short (*N*=5) and medium test (*N*=10) in all designs, whereas the SU calibration was the best model for the long test (*N*=20). An exception existed in the IM design with *ρ*=.9 and *N*=10, where the SU calibration was favored over the CM calibration. Again, it was not surprising to observe better model fit for the CM calibration, which incorporated more parameters than the other two calibrations. Interestingly, however, when the test was long enough (*N*=20), the SU calibration outperformed the CM calibration in terms of model fit. When a penalty of overparameterization was applied as in the BIC (see Table 21), the CU calibration was the best model for the short tests (*N*=5), the CM calibration the best model for the medium tests (*N*=10) with an exception for the PM design when *ρ*=.3 and *N*=10, and the SU calibration the best model for the long tests (*N*=20). This might suggest that the multidimensional complexity was not significant enough in the short test to require a multidimensional model, and the multidimensionality in the long test was strong enough to form independent clusters so that the SU calibration could effectively model such multidimensionality.

Table 20. Akaike Information Criterion

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.00	.00	.00	.97	1.00	.00	.03	.00	1.00
IM, ρ=.3	.00	.00	.00	1.00	1.00	.00	.00	.00	1.00
IM, ρ=.6	.00	.00	.00	1.00	.89	.00	.00	.11	1.00
IM, ρ=.9	.00	.00	.00	1.00	.35	.00	.00	.65	1.00
PM, ρ=.3	.00	.00	.00	.97	1.00	.00	.03	.00	1.00
PM, ρ=.6	.00	.00	.00	.97	1.00	.00	.03	.00	1.00
PM, ρ=.9	.00	.00	.00	.97	1.00	.00	.03	.00	1.00

Table 21. Bayesian Information Criterion

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.93	.13	.00	.07	.87	.00	.00	.00	1.00
IM, ρ=.3	1.00	.33	.00	.00	.67	.00	.00	.00	1.00
IM, ρ=.6	1.00	.22	.00	.00	.67	.00	.00	.11	1.00
IM, ρ=.9	1.00	.06	.00	.00	.35	.00	.00	.59	1.00
PM, ρ=.3	1.00	.93	.00	.00	.07	.00	.00	.00	1.00
PM, ρ=.6	1.00	.50	.00	.00	.50	.00	.00	.00	1.00
PM, ρ=.9	.97	.10	.00	.03	.90	.00	.00	.00	1.00

CHAPTER V

DISCUSSION

Optimal Test Design

A well-designed dual-purpose test was defined as one being able to produce scores that were reliable, accurate, and diagnostically useful. It was implied in Chapter IV that total score was accurately estimated in the UNI and IM designs, irrespective of the test length and calibration approaches, while in the PM designs, the total score was only accurately estimated in tests with long lengths or highly correlated true abilities (see Figure 8 to Figure 10). The total score was generally estimated with high reliability in all designs but the PM designs with short test length (N=5) and high degree of multidimensionality (ρ=.3 and ρ=.6), implying that when clear multidimensionality was present and the test was not long enough, the equally-weighted summation of, probably unreliable, subscores failed to provide a satisfactory estimation of the total score.

On the other hand, the subscores needed to be estimated with at least ten items per domain in order to achieve satisfactory accuracy in almost all designs except the concurrently calibrated PM designs with high degree of multidimensionality (ρ=.3 and ρ=.6; see Figure 11 to Figure 13). When the test development was restricted by the limited item pool, the IM designs was the most promising design in producing accurately estimated subscores for the short test. Within the IM designs, subscores were better estimated when a higher correlation was presented between true abilities, lending itself to borrowing more information across dimensions to improve estimation. This effect was absent in the PM designs, because the simple-structured item provided information only on its major dimension, leaving little information be borrowed across dimension. However, such results, namely score accuracy and reliability, should be interpreted in conjunction with the validity evidences, namely the (disattenuated) correlations between estimated subscores. Although the IM design allowed subscores to achieve high accuracy and reliability, the correlations of subscores substantially disagreed with their true values, meaning the underlying structure of subscores was not correctly recovered by either CM or SU calibration. Conversely, the PM design allowed the estimation of subscores to restore their true correlations, providing supportive evidence to the valid score use. Furthermore, subscores estimated with the traditional unidimensional IRT model were generally unlikely to provide additional diagnostic information to the total score. Regardless of test design, only subscores estimated with the multidimensional IRT model or separately with the unidimensional model were useful for the diagnosis purpose of the test (see Table 18).

In summary, the unidimensionally designed dual-purpose test could produce good scores, even for the short test with as few as five items per domain, using the traditional unidimensional calibration, but subscores failed to be supportive to the diagnostic function of the test. This was congruent with findings in the literature (Haberman, 2008; Sinharay et al., 2007; Sinharay, 2010). As Haberman (2008) pointed out, this was either due to low subscore reliability or high correlation between subscores. Since subscores in this case were estimated with high reliability (see Table 17), the high correlation was a more likely cause of the lack of diagnostic information (see Table 14). This also confirmed Luecht et al.’s (2006) and Wainer et al.’s (2001) assertion that items have to be multidimensional in order to realize the diagnostic function. Interestingly, however, when the unidimensional test was calibrated with the multidimensional or separate unidimensional approaches, diagnostic information was observed in subscores at the expense of score accuracy and reliability (see Figure 14). After considering the high disattenuated correlations between subscores, however, this did not justify the multidimensionally calibrated unidimensional test.

In the test with idiosyncratic multidimensionality and loosely correlated true abilities, scores were estimated with high accuracy and reliability using the CU calibration irrespective of test length, but when estimated using the CM and SU calibrations, satisfactory scores were obtainable only in the long tests. The quality of the unidimensional estimation of multidimensional responses was beyond expectation, because the unidimensional estimation was only expected to collect information on the reference composite dimension but ignore extra information in original dimensions. However, because the true domain-specific subscore was defined in this study as the summed ERS over items within the subtest, the ERS of each item, as well as the subscore, was already an aggregation of all information in all domains in the IM design even though it might have been dedicated to one particular content domain. Take an item with a=(.62, .16, .32, .21) and d=.30 for example. The item was coded as an ARI item, for which the a-parameter was greatest, but the ERS for this item would be computed as follows:

$$ERS = .1 + \frac{.9}{1+\exp[-1.7\times(.62\theta_1+.16\theta_2+.32\theta_3+.21\theta_4+.30)]} \tag{5.1}$$

This computation resulted in highly correlated subscores in the IM design. Highly correlated subscores when they were not expected to be so (e.g. ρ=.3 or .6), nevertheless, cast doubt on their validity. Although subscores were estimated with precision with the unidimensional IRT model, they failed to provide added value to the total score, and thus were useless for the diagnostic purpose. Diagnostically useful subscores were obtainable at the expense of score accuracy and reliability if multidimensional or separate unidimensional model was used for calibration and scoring. For a long test, the sacrifice might be acceptably small, yet for a short test, the sacrifice might be too remarkable to ignore.

Compared with the IM design, the corresponding PM design (ρ=.3) showed dramatic inaccuracy in subscores estimated with the unidimensional model. Since the CU calibration failed to produce diagnostically informative subscores, the dramatic drop of score accuracy was not concerned. For the CM and SU calibrations, the accuracy and reliability were slightly smaller in the PM design than in the IM design, because of higher degree of multidimensionality. That is, even though the domain-specific abilities were correlated at the same degree in the IM and PM designs, less amount of information could be borrowed across dimension in the PM design than in the IM design due to the simple-structure in the PM design, and estimated subscores were consequently even more loosely correlated in the PM design than subscores in the IM design (see Table 14). It is, however, worth noting that the underlying structure of subscores was well restored in the test with PM design. Thus, the slight lower accuracy and reliability compensated for the validity and rigor of interpretability and usability of subscores.

When the true abilities were correlated with one another at a higher degree, accuracy and reliability of subscores in the test of the IM design would increase, especially in the short test. This was because higher correlation allowed more information to be borrowed across domains to improve the estimation. It confirmed de la Torre et al.'s (2011) finding that the correlation-based information borrowing mechanism was most effective for highly-correlated short-length subtests. This effect of information borrowing was in absence in the PM design, due to the simple-structure constraint. Again, the accuracy and reliability declined without the borrowing of information, but the rigor of score interpretability increased in return.

Calibration

The effect of the calibration approach on the estimation of the total score was almost negligible across various test designs. In the UNI and IM designs, the total score was accurately estimated with all three calibration approaches, regardless of the test length. While this finding met the expectation for the unidimensional calibration (e.g., CU) of unidimensional test (e.g., UNI) and the multidimensional calibration (e.g., CM and SU) of multidimensional test (e.g., IM), the good estimation results of the total score in the crossed conditions (i.e., the multidimensional calibration of unidimensional test and the unidimensional calibration of multidimensional test) was also not entirely unexpected. Extraordinary high ability correlation estimated in the multidimensional calibration (see Table 9) might explain why the multidimensional calibration was still good for the unidimensional test, and highly correlated true scores in the multidimensional test (see Table 7) might explain why the unidimensional calibration was sufficient for the multidimensional test.

The three calibration approaches functioned very comparably in the PM designs in terms of the totals score too. However, unlike in the UNI and IM designs, the estimation quality was a function of test length and true ability correlations. This was because the PM deigns was a "truly" multidimensional design in the sense that an item, by design, solely provided measurement information in the major dimension, leaving other three minor dimensions blank. Consequently, when the subscores were inaccurately and unreliably estimated with limited items in subtests, the total score, which simply summed up distinct subscores, was not expected to be a good estimator of the true total score.

When subscores were highly correlated in the UNI and IM designs, the unidimensional calibration proved to be effective in estimating subscores. Yet in the PM designs where subscores were relatively distinct, the item characteristics on the reference composite dimension capture by the CU calibration failed to provide good estimations for subscores. On the other hand, subscores were accurately estimated with the CM and SU calibrations for the long test, but just decently for the short test in all designs. Furthermore, a subtle disparity between the CM and SU calibrations appeared in the IM designs. When items were complex-structured in IM designs, the CM calibration exploited that information, borrowing it across dimension to improve subscore estimation and resulting in better subscores than the SU calibration. When subscores were simple-structured in PM designs, the CM and SU calibrations were nearly identical.

Both the CM and SU calibration proved to fit the response data better than the CU calibration in all test designs, as AIC statistics implied (see Table 19). It was expected so for the CM calibration, which incorporated more parameters and ought to yield better model fit. The SU calibration used equal number of parameters with the CU calibration, yet it still outperformed the CU calibration in terms of model fit, possibly because the separate unidimensional calibration approach gave the model more flexibility in fitting the response data. An interesting pattern was implied by BIC statistics, after penalizing models for overparameterization: The CU calibration generally fit the response data in the short test best and the SU calibration fit the long test best. It might suggest that the multidimensionality was too weak to ask for a multidimensional model in the short test and strong enough to form multiple independent clusters in the long test.

In practice, some practitioners analyzed the same response data set with both unidimensional and multidimensional approaches to realize the dual purposes of the test, for which the unidimensional analysis produced the total score and the multidimensional analysis subscores (Brandt, 2008, 2010). In spite of the difficulty of explaining the methodology and results of this practice to the public, results in this study might justify that this practice was psychometrically sound. That is, both the total score produced from the CU calibration and subscores produced from the CM or SU calibration were accurate and reliable. More important, the total score and subscores were on the same raw score scale when the ERS was used.

Alternative Scoring Methods

Generally, alternative scoring methods were either unnecessary or ineffective. The scoring methods based on Kelley's true score regression (i.e., Longford's composite total score and Wainer et al.'s augmented scores) malfunctioned in the CU calibration, in which unidimensional subscores were highly correlated as they should and thus confused the alternative scoring methods with multicollinearity (see Figure 14 to Figure 20). With the CM and SU calibrations, the scoring methods for the total score were almost identical to the raw total scores, probably because the total score was too good to be improved. The augmented subscores provided some improvements to raw subscores, especially for the test with few items and high dimensional correlation, as found in de la Torre et al.'s (2011) study. The improvements, however, were practically trivial. The results also implied that the score augmentation was ineffective when scores were already very reliable and accurate to their true values.

CHAPTER VI

CONCLUSIONS

Inspired by the increasing demands for psychometrically sound dual-purpose tests in practice, the present study explored how to design and analyze the dual-purpose test through systematical comparisons of various plausible test developmental and analytic options. In order to obtain accurate and reliable total score and subscores from the test, the test should be designed with multidimensionality and at least 10 items per domain and be analyzed with the multidimensional or separate unidimensional IRT model. Specifically, the unidimensional dual-purpose test was able to produce reliable and accuracy but not diagnostically meaningful scores. Subscores obtained from an essentially unidimensional test were either unable to provide added value to the total score according to the PRMSE criterion or homogeneous to each other according to disattenuated correlations. The idiosyncratic multidimensional design was able to yield accurate, reliable, and diagnostically useful scores, but the validity of the diagnostic subscores was questionable, whose correlation disagreed with the true correlational structure. Consequently, even though subscores were identified distinct from the total score according to the PRMSE criterion, they were still nearly identical to each other according to the disattenuated correlations. On the other hand, the principled multidimensional design showed slightly lower accuracy and reliability in scores due to principled “simple structure” of test design, but this sacrifice of accuracy and reliability ensured the interpretability and validity of diagnostic subscores, whose empirical correlational structure approximated the true structure.

Furthermore, with respect to the calibration methods, the unidimensional calibration was found failing to distinguish subscores, and thus failing to give subscores useful diagnostic information, even though the subscores sometimes appeared more accurate and reliable than subscores obtained with the other two calibrations. The confirmatory multidimensional calibration and separate unidimensional calibration delivered very comparable results, confirming previous findings that suggested the plausibility of analyzing multidimensional response data using unidimensional IRT approach (Luecht & Miller, 1992; Luecht et al., 2006). Alternative scoring methods were found either inappropriate to use or offering insignificant improvements over the raw scores.

There were three major methodological limitations in this study. The first limitation concerned the definition of true scores in multidimensional designs. In multidimensional designs, true scores were the expected raw scores aggregated over the subtest or the complete test. This avoided the conversion between subscores and total scores on the θ scale, but whether this true score mirrored the ability of true latent trait was questionable. Each item's expected raw score, namely the probability of getting a correct response given by the M3PL model, was an aggregated score, accounting for information on all dimensions. As a result, true subscores were correlated with one another at a high degree, even though true latent abilities were correlated at a far lower degree. In this sense, results of this study indicated how adequately different test designs recovered true “scores” instead of true abilities. However, it should be remembered that using the true latent ability for comparison would have avoided this discrepancy between scores and abilities but encountered the conversion problem, for which there had not been a perfect solution. Moreover, this scoring method used in this study, though not flawless, resembled the scoring practice used in most test batteries or tests with subtests so that results obtained in this study might be illuminating to practitioners.

Next, the complexity of multidimensionality was not manipulated in the idiosyncratic multidimensional design, which would have shown an impact on the results. Measurement information was randomly allotted into minor assessment dimensions in the IM designs. This treatment made the information borrowing mechanism ambiguous. Should item development strive to leave minimal information in minor assessment dimensions or evenly allot information to both major and minor assessment dimensions? Question of this kind was not addressed, but essential to interpretations of findings in this study.

Lastly, as a common limitation to virtually all simulation studies, this study was limited by the data generation mechanism. While the simulation was an efficient and economic research method, it never assured the generalizability of results and findings obtained from the simulated data. Thus, findings of this study would be greatly supplemented by evidences obtained from analyses using real data.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*: Brooks/Cole Monterey, CA.

American Educational Research, A., American Psychological, A., & National Council on Measurement in, E. (1999). *Standards for educational and psychological testing*.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*(1)*,* 37-48.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. *IERI monograph series: Issues and methodologies in large-scale assessments, 1*, 51-69.

Brandt, S. (2010). Estimating tests including subtests. *Journal of applied measurement, 11*(4), 352.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620-639.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Fraser, C. (1993). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory (Version 2). Armidale. *New South Wales, Australia: The University of New England Center for Behavioral Studies*.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.

Graybill, F. A., & Deal, R. B. (1959). Combining unbiased estimators. *Biometrics, 15*(4), 543-550.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229.

Haertel, E. H. (2006). Reliability. In R. L. Brennen (Ed.), *Educational Measurement* (4 ed., Vol. 4, pp. 65-110). Westport, CT: Praeger Publishers.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications* (Vol. 7): Springer.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*(4), 358-377.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Kane, M. T. (2006). Validation. *Educational measurement, 4,* 17-64.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Longford, N. T. (1997). Shrinkage estimation of linear combinations of true scores. *Psychometrika, 62*(2), 237-244.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: ERIC.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*(4), 389-404.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*(3), 279-293.

McDonald, R. P. (1982). Linear Versus Models in Item Response Theory. *Applied Psychological Measurement, 6*(4), 379-396.

McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference*. Minneapolis, MN: University of Minnesota, Department of Psychology.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items or testlets scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439-483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379-416.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*(4), 661-679.

Nandakumar, R. (2005). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*(2), 99-117.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics, 18*(1), 41-68.

Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: Diagnosis and implications. *Journal of Educational Measurement, 45*(3), 201-223.

Puhan, G., Sinharay, S., Haberman, S., & 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.

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*: Springer.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193-203.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361-373.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331-354.

Segall, D. O. (2010). Principles of multidimensional adaptive testing *Elements of adaptive testing* (pp. 57-75): Springer.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150-174.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293-325.

Stout, W. F. (2005). DIMTEST (Version 2.0)[Computer software]. *Champaign, IL: The William Stout Institute for Measurement*.

Swygert, K. A., McLeod, L. D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 217-250). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.)*, Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis, MN: University of Minnesota, Department of Psychology.

Wainer, H., Bradlow, E. T., & Du, Z. (2002). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. *Computerized adaptive testing: Theory and practice*, 245-269.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement, 24*(3), 185-201.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., . . . Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. *Test scoring*, 343-387.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Wilson, D. T., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*: SSI, Scientific Software International.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2006). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291-309.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. *Chicago: Scientific Software*.

APPENDIX A

GENERATION OF MULTIDIMENSIONAL ITEM PARAMETERS

The probability of getting a correct response to an item in the unidimensional 3PL model was given by

$$P\left(u_{ij} = 1|\theta_i, a_j, b_j, c_j\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7a_j\left(\theta_i - b_j\right)\right]}$$

Generalizing of the unidimensional 3PL model to the compensatory multidimensional 3PL model, the probability of getting a correct response to an *m*-dimensional item was given by

$$P\left(u_{ij} = 1|\theta_i, a_j, b_j, c_j\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7\sum_{k=1}^{m} a_{jk}\left(\theta_{ik} - b_{jk}\right)\right]}$$

Multiplying through the term $a(\theta - b)$ by *a*-parameter and replacing $-ab$ with *d*, the above equation was identical to

$$P\left(u_{ij} = 1|\theta_i, a_j, b_j, c_j\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7\left(\sum_{k=1}^{m} a_{jk}\theta_{ik} + d_j\right)\right]}$$

where

$$d_j = \sum_{k=1}^{m} a_{jk}b_{jk}$$

The above equation was often called the slope/intercept form of the IRT model, which reduced the number of item parameters from *2m* to *m+1*.

Since the *d*-parameter was a derivative from dimensional *a*- and *b*-parameters, its distribution is unclear. This posed a question to the generation of multidimensional item parameters: How should *d*-parameters be generated? By theory, its distribution was a complex mixture of lognormal and normal distributions, On the other hand, it might be reasonable to assume a normal distribution for *d*-parameters, assuming it as a multidimensional difficulty parameter that is analogous to the unidimensional difficulty parameter. Therefore, a small simulation study was conducted to compare two item parameter generation methods. Both methods generated *a*-parameters from the lognormal distribution. Yet, the first method directly generated *d*-parameters from the normal distribution, whereas the second method generated dimensional *b*-parameters and converted them to *d*-parameters. Generated item parameters were used to replicate the rest of the study described in this study with one simulation replication. Results regarding score estimation were presented in Table 21 and Table 22. Statistics in these two tables were almost identical, meaning these two methods of generating multidimensional item parameters were equivalent.

Table 22. Correlations between True and Estimated Scores when d-parameters were Directly Generated

		CU Calibration					CM Calibration					CU Calibration				
			Subscore					Subscore					Subscore			
Design	N	Total	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4
UNI	5	.93	.92	.93	.93	.93	.92	.75	.78	.77	.76	.92	.76	.78	.77	.76
UNI	10	.96	.96	.96	.96	.96	.96	.88	.86	.87	.86	.96	.88	.86	.87	.85
UNI	20	.98	.98	.98	.98	.98	.98	.93	.93	.93	.92	.98	.93	.93	.93	.93
IM, *ρ*=.3	5	.96	.93	.94	.93	.94	.95	.86	.87	.83	.82	.95	.86	.86	.83	.81
IM, *ρ*=.3	10	.97	.95	.93	.95	.95	.97	.92	.90	.91	.91	.97	.92	.90	.90	.91
IM, *ρ*=.3	20	.98	.96	.97	.95	.96	.98	.94	.94	.94	.94	.99	.95	.95	.95	.95
IM, *ρ*=.6	5	.96	.95	.96	.95	.95	.96	.88	.89	.86	.85	.95	.87	.88	.85	.84
IM, *ρ*=.6	10	.98	.96	.96	.97	.97	.98	.93	.92	.93	.91	.98	.94	.92	.93	.93
IM, *ρ*=.6	20	.99	.98	.98	.97	.97	.98	.95	.95	.94	.94	.99	.96	.96	.96	.96
IM, *ρ*=.9	5	.97	.96	.97	.97	.97	.96	.89	.90	.88	.86	.96	.88	.89	.88	.86
IM, *ρ*=.9	10	.98	.97	.98	.98	.98	.98	.93	.93	.93	.92	.98	.94	.93	.93	.94
IM, *ρ*=.9	20	.98	.98	.98	.98	.98	.98	.95	.95	.95	.94	.99	.97	.97	.97	.97
PM, *ρ*=.3	5	.88	.60	.64	.54	.61	.87	.80	.82	.79	.78	.87	.80	.82	.79	.78
PM, *ρ*=.3	10	.92	.57	.63	.72	.62	.93	.87	.87	.89	.88	.92	.87	.88	.89	.88
PM, *ρ*=.3	20	.96	.65	.65	.65	.68	.96	.92	.93	.93	.93	.96	.92	.93	.93	.93
PM, *ρ*=.6	5	.91	.76	.78	.72	.77	.91	.81	.81	.79	.79	.91	.81	.81	.78	.79
PM, *ρ*=.6	10	.95	.78	.80	.81	.81	.95	.87	.88	.89	.89	.95	.87	.88	.89	.88
PM, *ρ*=.6	20	.97	.81	.81	.82	.83	.97	.92	.93	.93	.93	.97	.93	.93	.93	.93
PM, *ρ*=.9	5	.93	.89	.90	.89	.90	.93	.81	.81	.80	.79	.93	.80	.81	.79	.79
PM, *ρ*=.9	10	.96	.91	.92	.93	.92	.96	.86	.88	.89	.88	.96	.86	.88	.89	.88
PM, *ρ*=.9	20	.98	.94	.94	.94	.94	.98	.92	.93	.93	.93	.98	.92	.93	.93	.93

Table 23. Correlations between True and Estimated Scores when d-parameters were Indirectly Generated

		CU Calibration					CM Calibration					CU Calibration				
			Subscore					Subscore					Subscore			
Design	N	Total	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4
UNI	5	.93	.92	.92	.93	.92	.92	.78	.80	.75	.74	.92	.78	.80	.75	.74
UNI	10	.97	.96	.96	.96	.97	.96	.88	.88	.88	.88	.96	.88	.88	.88	.88
UNI	20	.98	.98	.98	.98	.98	.98	.94	.93	.92	.92	.98	.94	.94	.92	.92
IM, *ρ*=.3	5	.94	.91	.91	.92	.89	.94	.82	.83	.80	.82	.94	.82	.83	.80	.82
IM, *ρ*=.3	10	.98	.96	.96	.95	.95	.97	.91	.90	.91	.91	.97	.92	.90	.91	.91
IM, *ρ*=.3	20	.98	.97	.97	.96	.97	.98	.95	.94	.94	.94	.99	.96	.95	.95	.95
IM, *ρ*=.6	5	.96	.94	.94	.95	.93	.95	.85	.85	.84	.85	.95	.84	.85	.83	.84
IM, *ρ*=.6	10	.98	.97	.97	.97	.97	.98	.93	.92	.93	.92	.98	.94	.92	.93	.93
IM, *ρ*=.6	20	.98	.98	.98	.97	.98	.98	.95	.95	.95	.95	.99	.96	.96	.96	.96
IM, *ρ*=.9	5	.96	.96	.96	.96	.96	.96	.87	.87	.86	.87	.96	.86	.87	.86	.87
IM, *ρ*=.9	10	.98	.97	.98	.98	.97	.98	.93	.92	.92	.92	.98	.94	.93	.93	.93
IM, *ρ*=.9	20	.99	.98	.98	.98	.98	.98	.95	.95	.95	.95	.99	.97	.96	.97	.96
PM, *ρ*=.3	5	.85	.58	.68	.52	.56	.85	.77	.79	.74	.76	.85	.77	.79	.74	.76
PM, *ρ*=.3	10	.93	.66	.64	.63	.61	.93	.88	.87	.87	.87	.93	.88	.87	.87	.87
PM, *ρ*=.3	20	.95	.74	.58	.70	.60	.96	.94	.91	.93	.92	.96	.94	.92	.93	.92
PM, *ρ*=.6	5	.90	.74	.78	.72	.74	.89	.74	.80	.75	.77	.89	.75	.80	.74	.77
PM, *ρ*=.6	10	.95	.81	.79	.79	.78	.95	.88	.87	.87	.86	.95	.88	.87	.87	.86
PM, *ρ*=.6	20	.97	.83	.81	.82	.78	.97	.93	.93	.92	.91	.97	.93	.93	.93	.91
PM, *ρ*=.9	5	.92	.88	.89	.88	.88	.91	.76	.80	.74	.76	.91	.76	.80	.73	.76
PM, *ρ*=.9	10	.95	.92	.91	.91	.92	.95	.87	.84	.84	.86	.95	.87	.84	.84	.86
PM, *ρ*=.9	20	.98	.94	.94	.94	.93	.98	.93	.93	.92	.91	.98	.93	.93	.93	.91

APPENDIX B

TRUE SCORE DISTRIBUTIONS

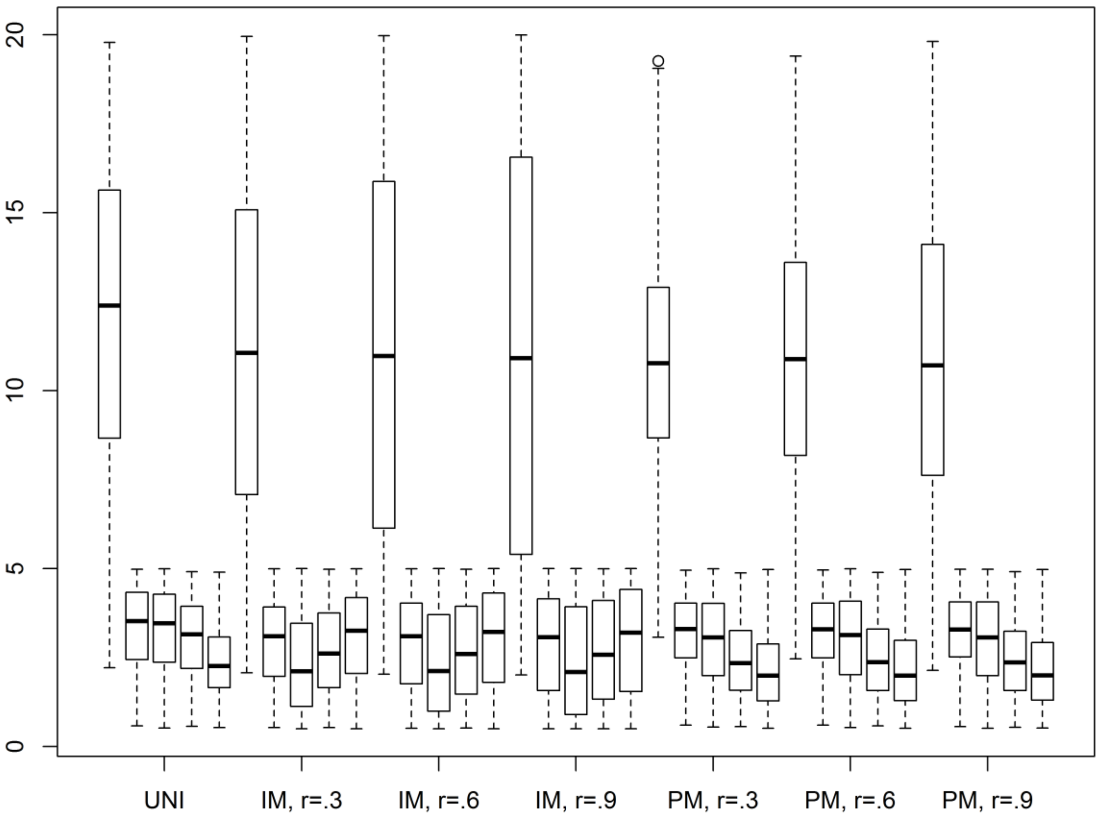

Figure 21. True Score Distributions when *N*=5

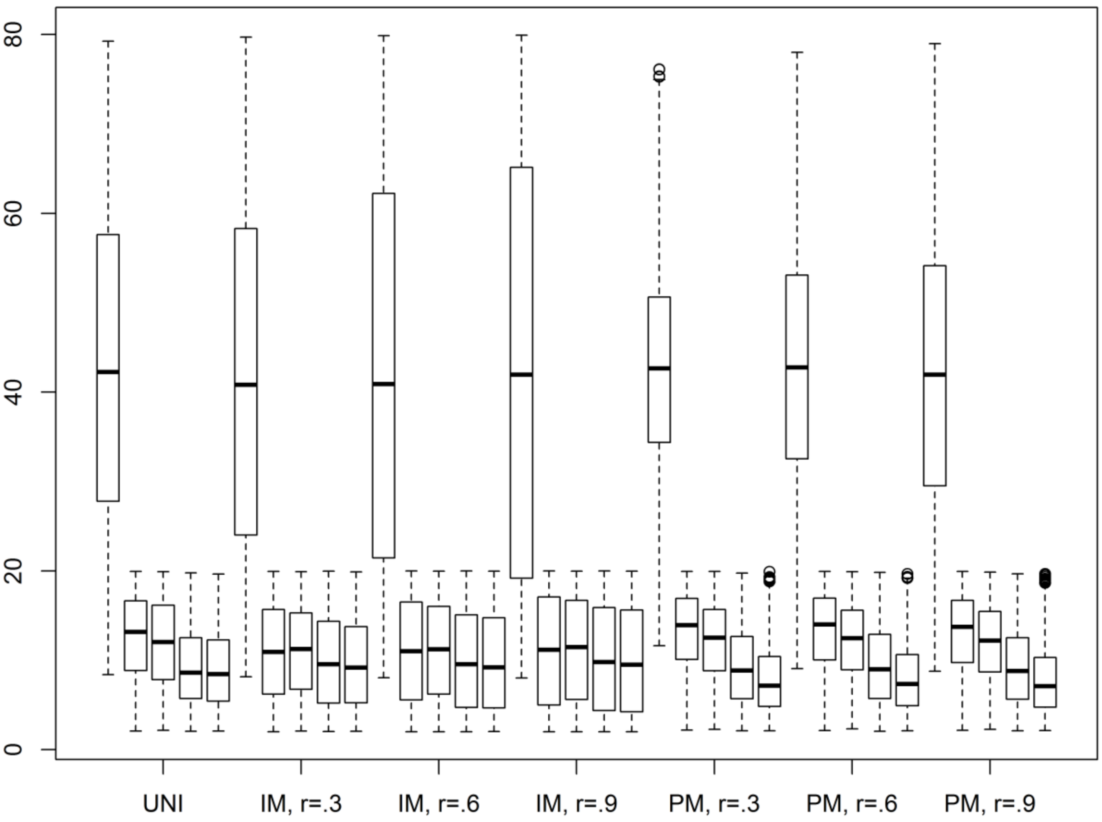

Figure 22. True Score Distributions when *N*=20

APPENDIX C

ESTIMATED SCORE DISTRIBUTIONS

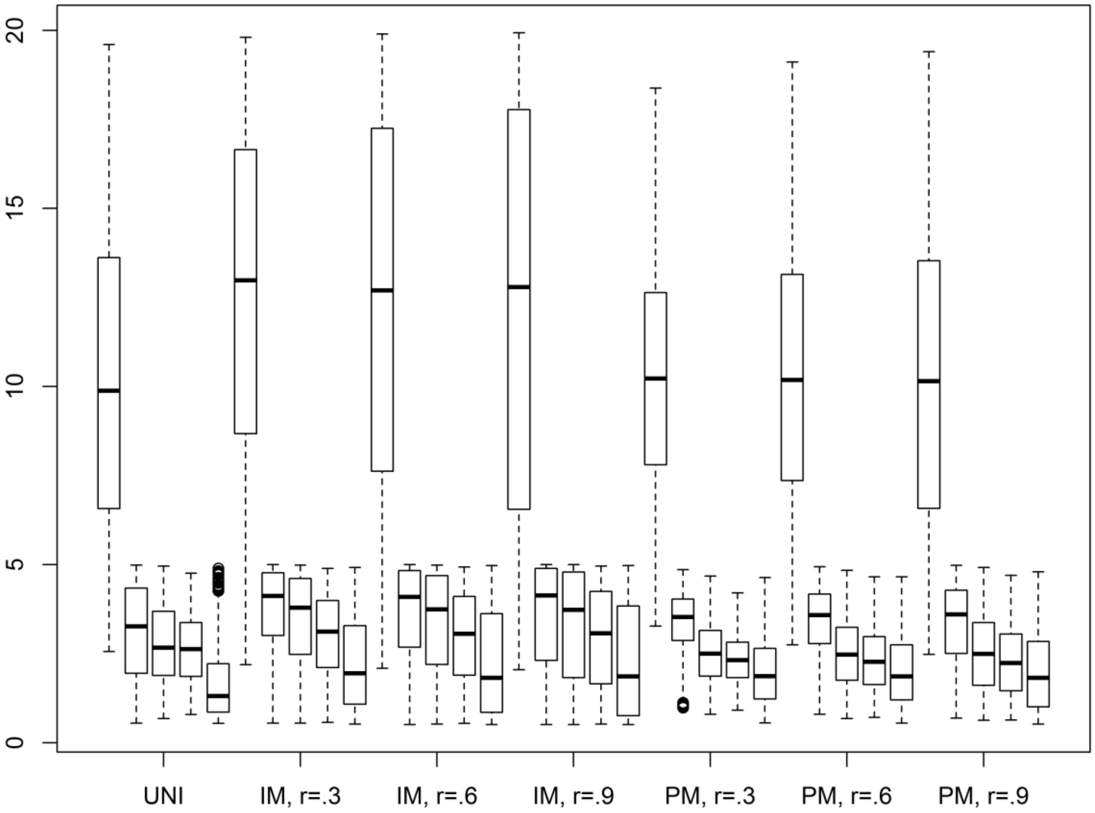

Figure 23. Distributions of Raw Scores Estimated with the CU Calibration when *N*=5

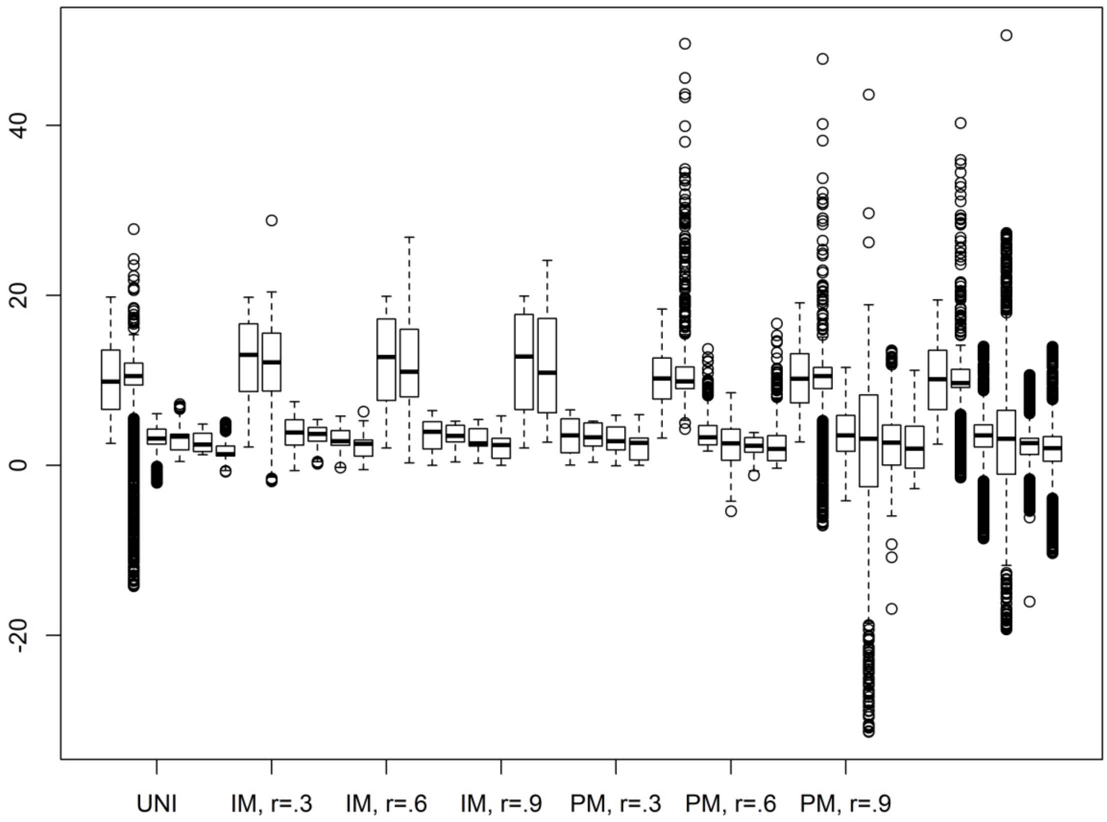

Figure 24. Distributions of Alternative Scores Estimated with the CU Calibration when *N*=5

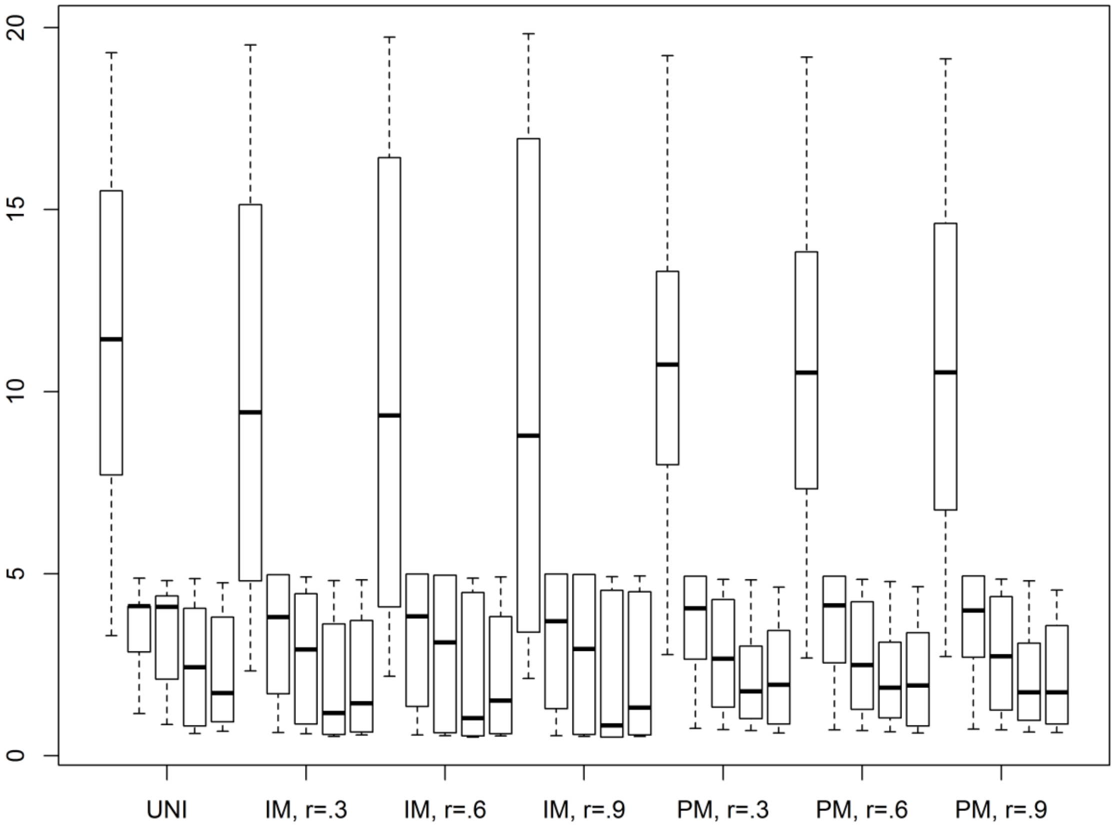

Figure 25. Distributions of Raw Scores Estimated with the CM Calibration when *N*=5

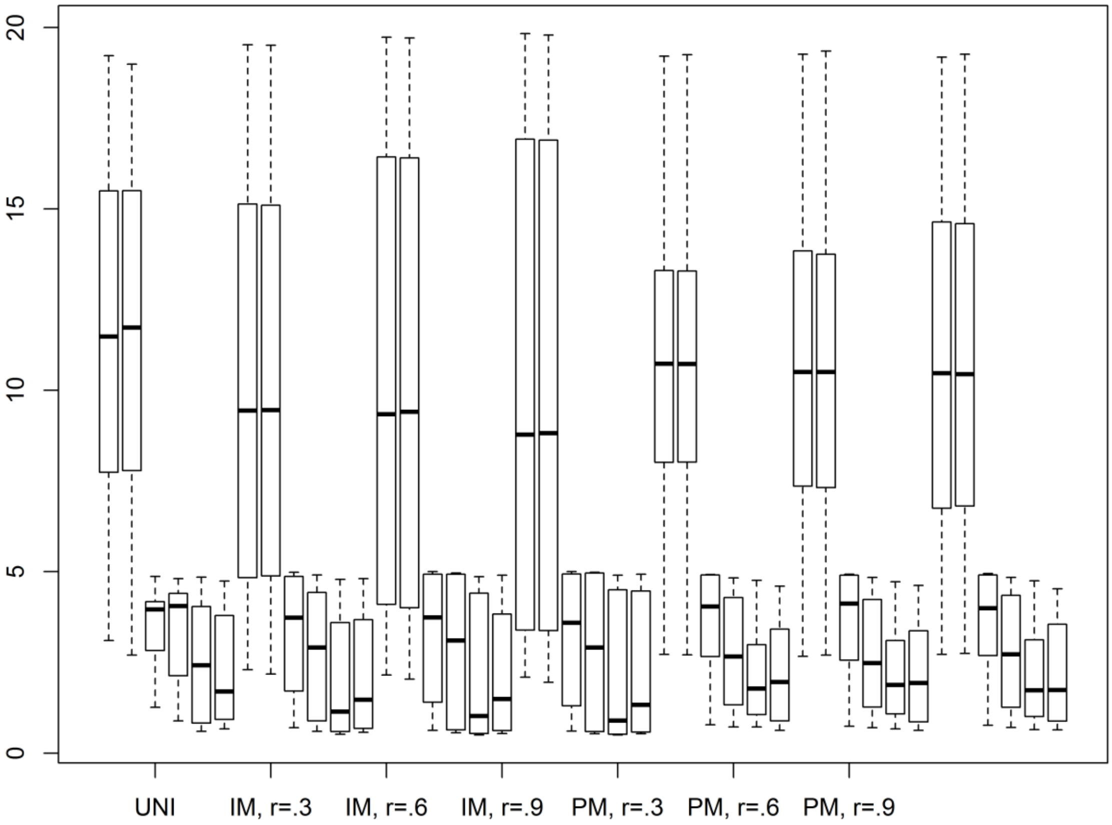

Figure 26. Distributions of Alternative Scores Estimated with the CM Calibration when *N*=5

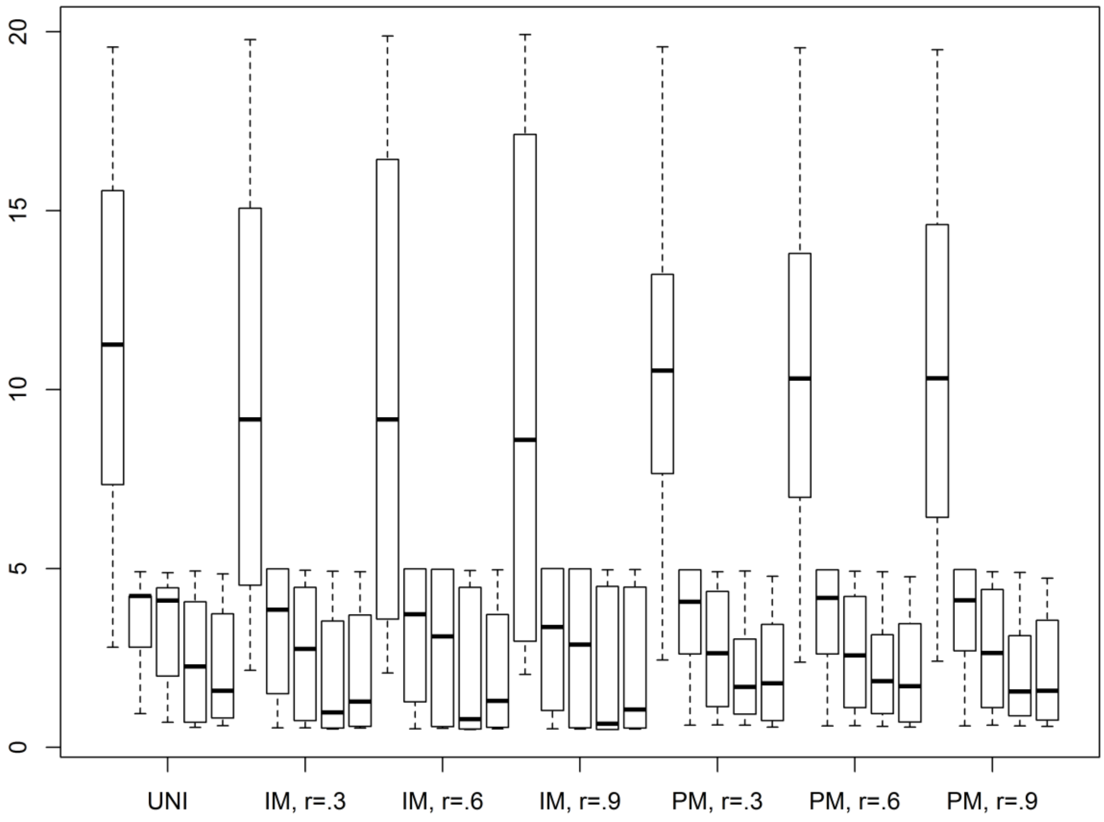

Figure 27. Distributions of Raw Scores Estimated with the SU Calibration when *N*=5

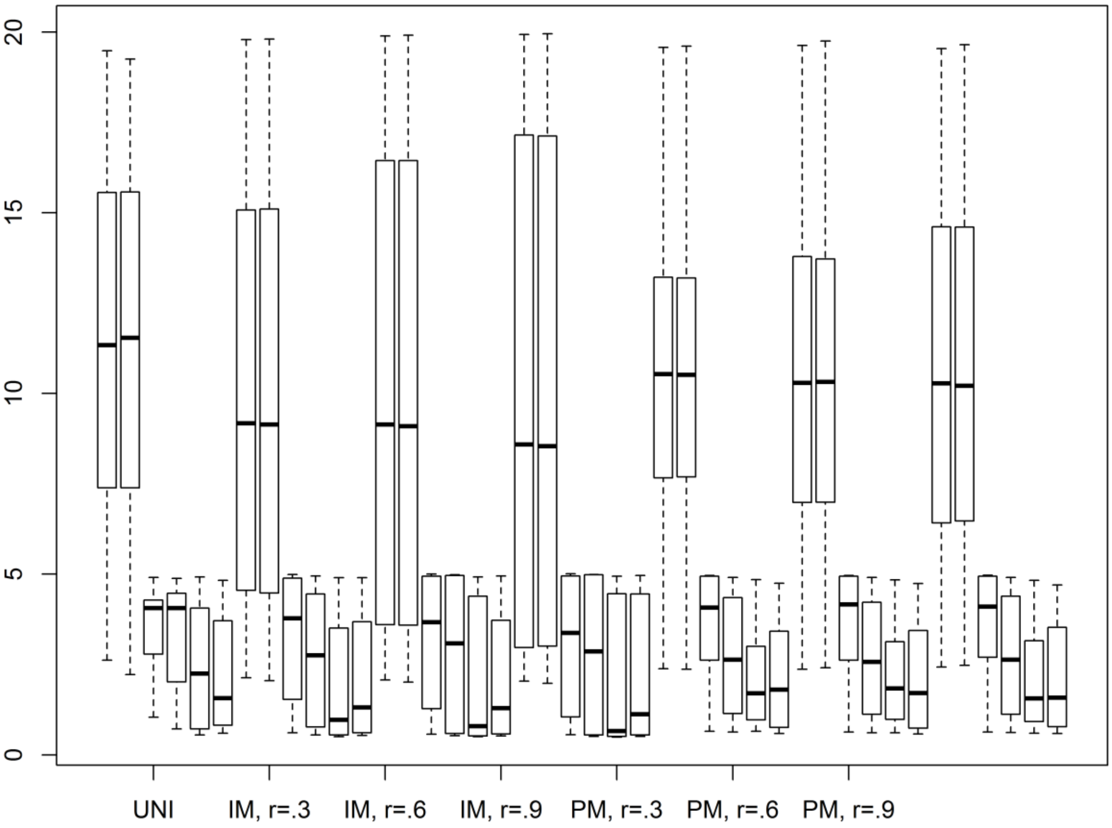

Figure 28. Distributions of Alternative Scores Estimated with the SU Calibration when *N*=5

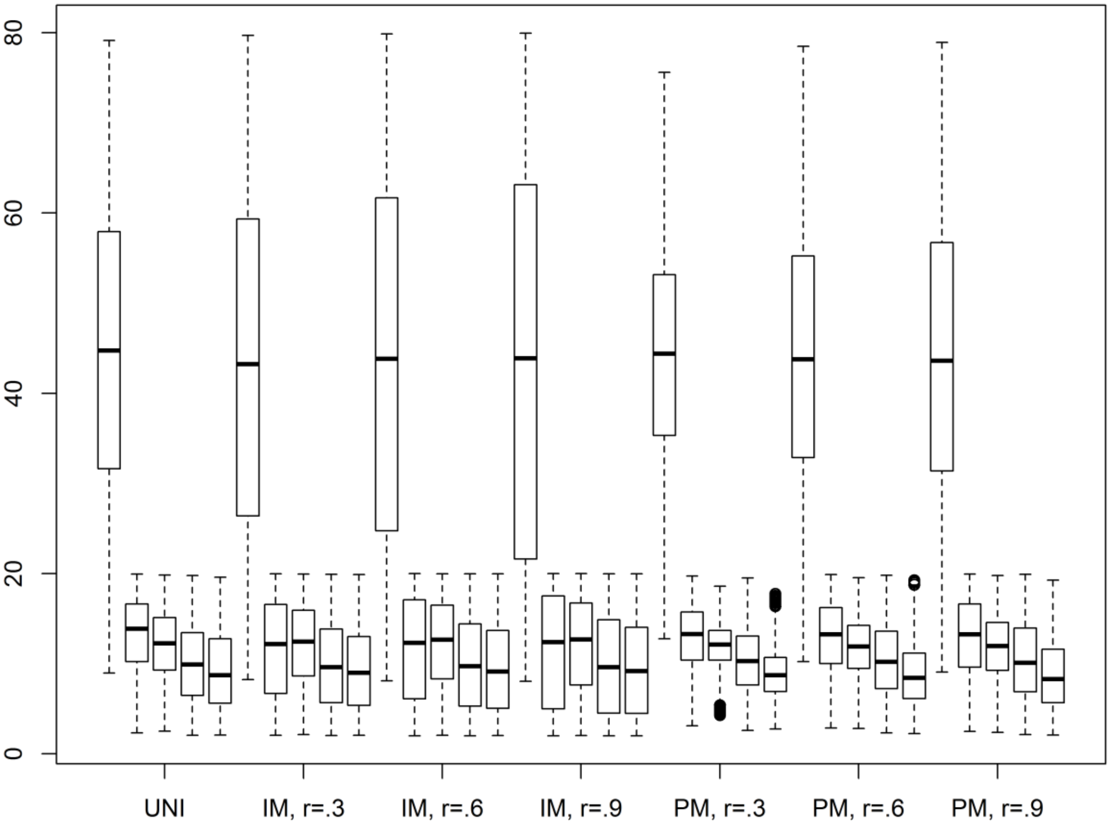

Figure 29. Distributions of Raw Scores Estimated with the CU Calibration when *N*=20

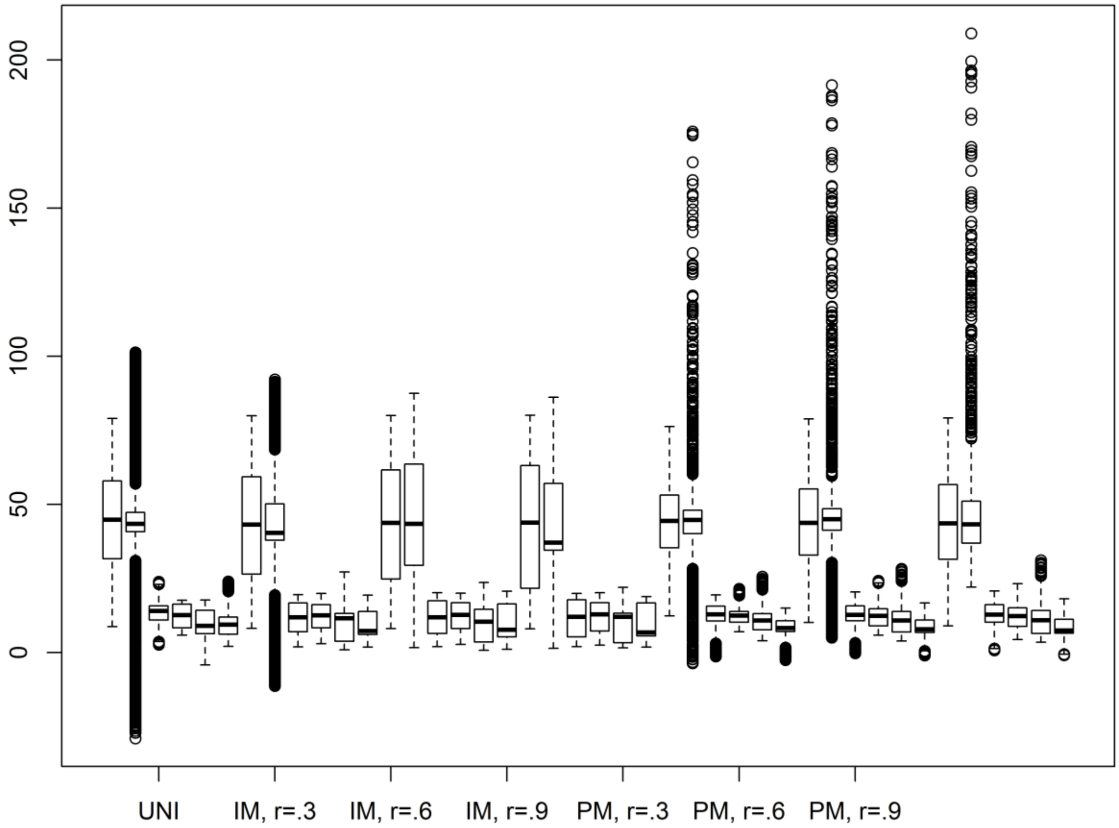

Figure 30. Distributions of Alternative Scores Estimated with the CU Calibration when *N*=20

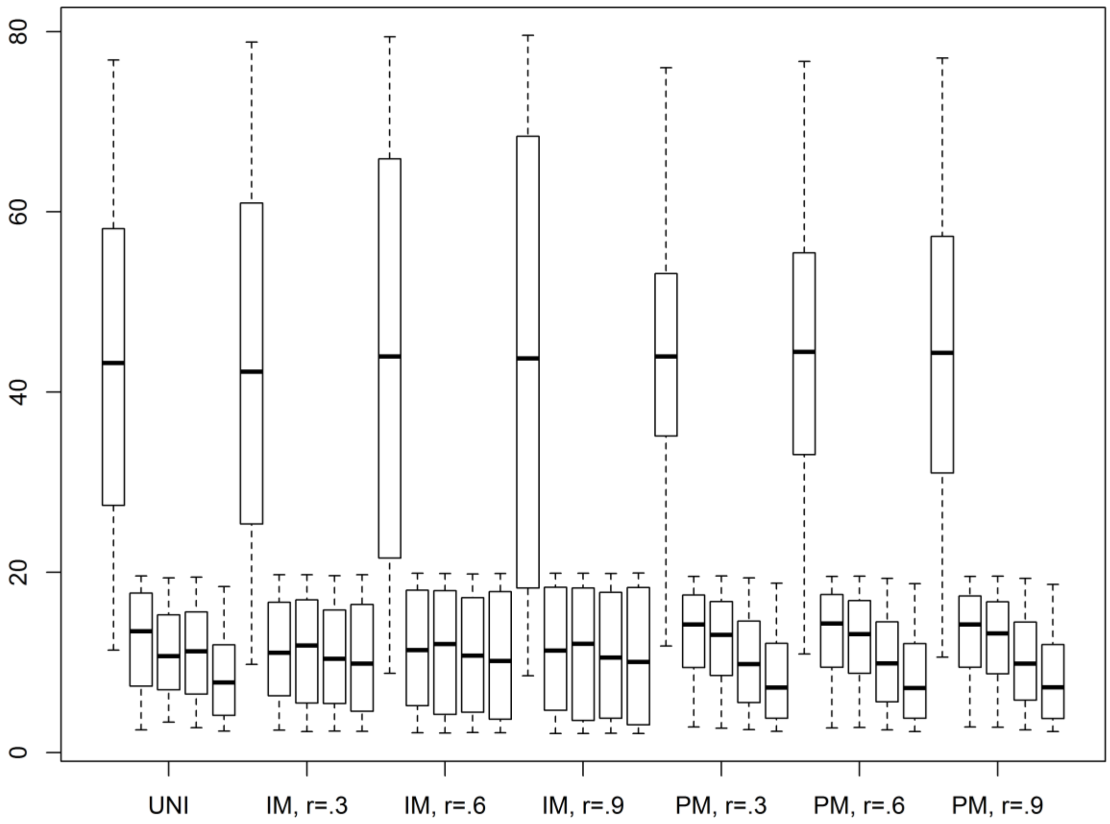

Figure 31. Distributions of Raw Scores Estimated with the CM Calibration when *N*=20

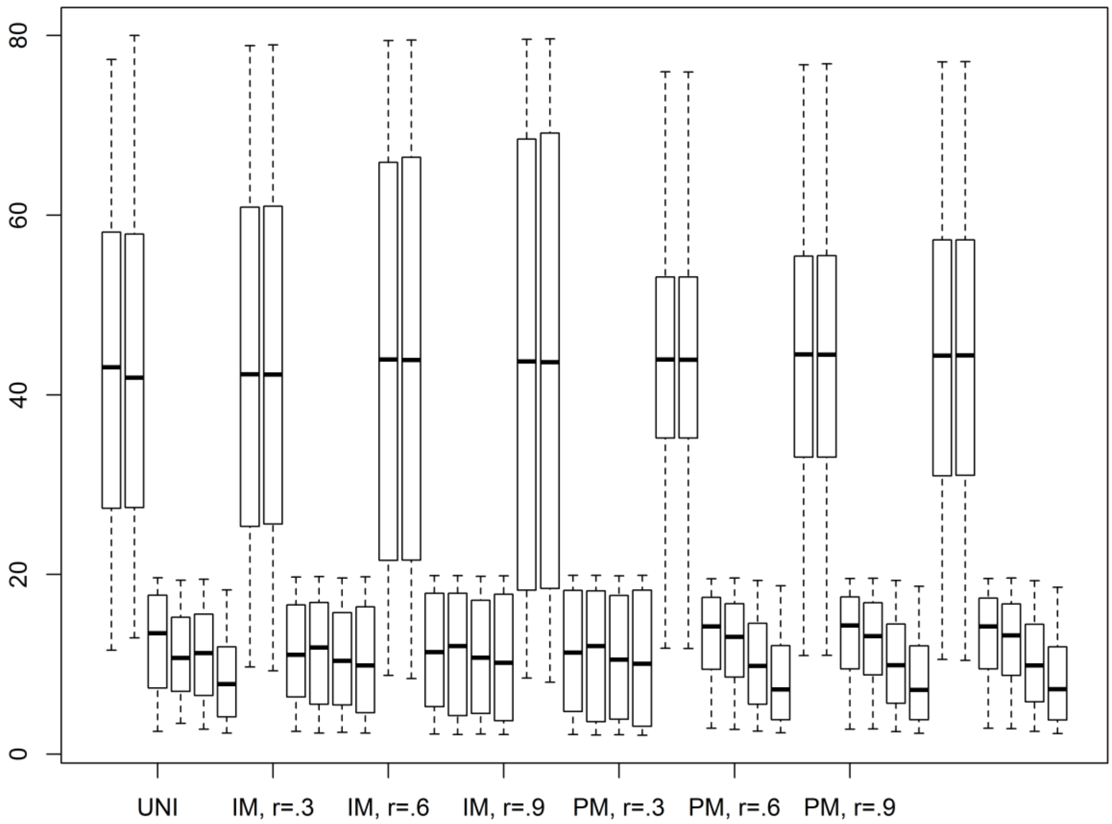

Figure 32. Distributions of Alternative Scores Estimated with the CM Calibration when *N*=20

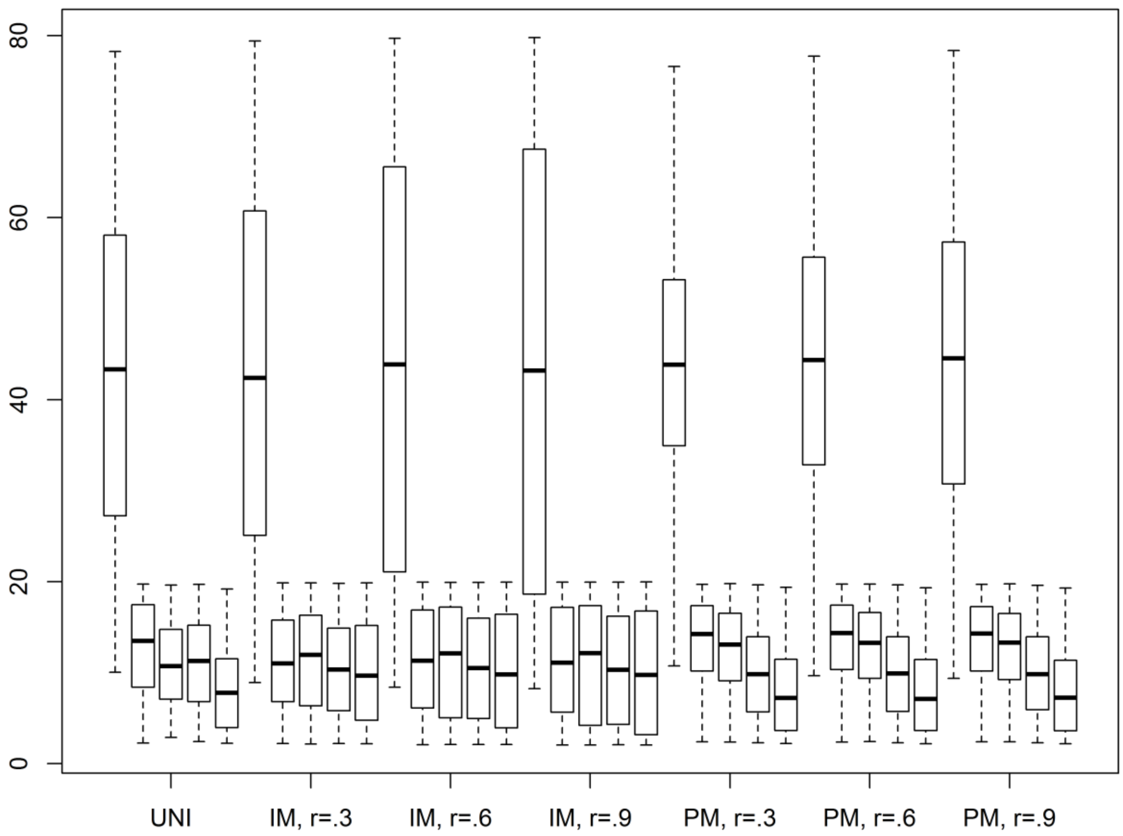

Figure 33. Distributions of Raw Scores Estimated with the SU Calibration when *N*=20

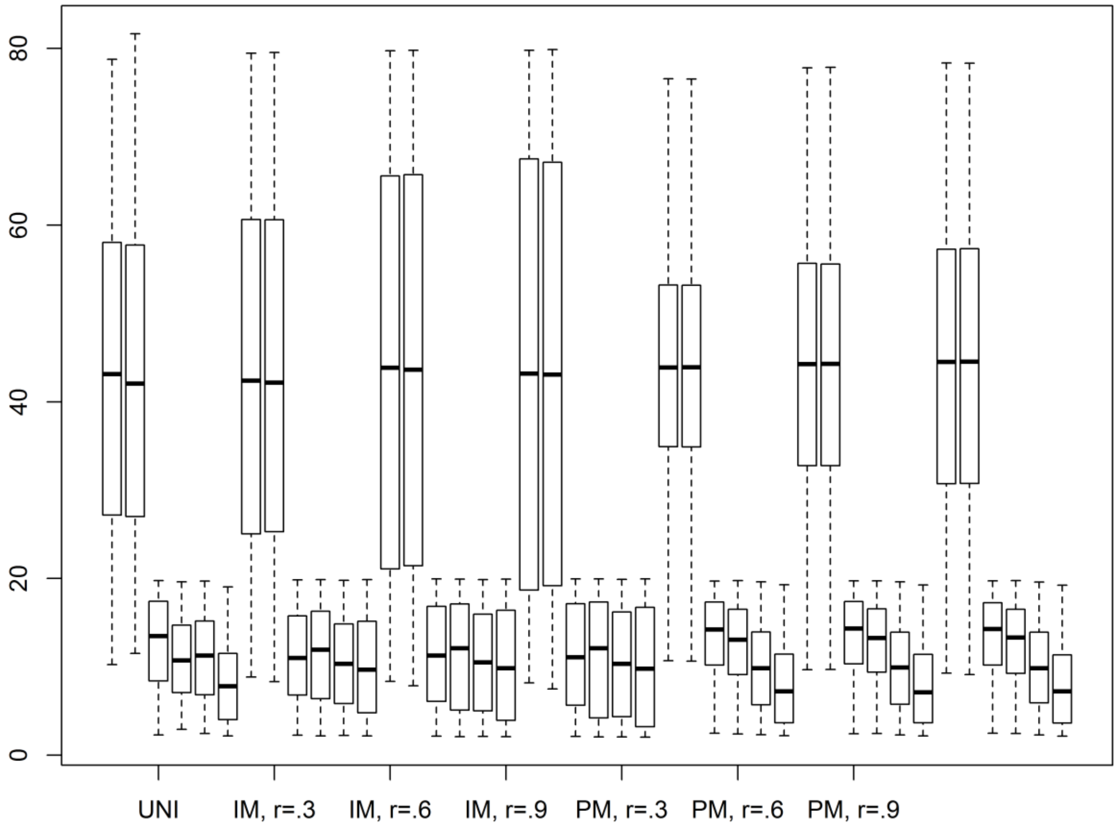

Figure 34. Distributions of Alternative Scores Estimated with the SU Calibration when *N*=20

APPENDIX D

SUPPORTING COMPARISON RESULTS

Table 24. Correlations between True and Estimated Raw Total Scores

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.924	.962	.979	.918	.959	.979	.917	.959	.979
IM, r=.3	.945	.972	.984	.940	.970	.983	.938	.970	.985
IM, r=.6	.958	.978	.985	.952	.976	.985	.949	.976	.989
IM, r=.9	.965	.980	.985	.959	.979	.984	.956	.979	.990
PM, r=.3	.849	.917	.956	.850	.920	.957	.849	.920	.959
PM, r=.6	.893	.944	.971	.889	.943	.970	.887	.943	.971
PM, r=.9	.917	.957	.978	.912	.955	.977	.910	.955	.978

Table 25. Correlations between True and Estimated Raw Subscore 1 (ARI)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.916	.958	.977	.772	.872	.927	.771	.871	.930
IM, r=.3	.919	.949	.962	.825	.904	.942	.822	.905	.950
IM, r=.6	.944	.967	.976	.856	.921	.948	.852	.924	.961
IM, r=.9	.960	.977	.983	.875	.929	.950	.870	.933	.966
PM, r=.3	.583	.649	.664	.766	.870	.926	.765	.871	.929
PM, r=.6	.747	.798	.814	.765	.871	.927	.764	.871	.930
PM, r=.9	.875	.918	.939	.768	.869	.927	.767	.869	.930

Table 26. Correlations between True and Estimated Raw Subscore 2 (ALG)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.921	.961	.979	.774	.868	.928	.772	.867	.931
IM, r=.3	.922	.948	.960	.824	.900	.942	.821	.900	.949
IM, r=.6	.947	.967	.975	.854	.918	.948	.849	.919	.960
IM, r=.9	.962	.977	.983	.873	.927	.950	.867	.930	.966
PM, r=.3	.603	.627	.666	.773	.865	.926	.772	.865	.929
PM, r=.6	.749	.786	.813	.773	.864	.927	.771	.864	.930
PM, r=.9	.881	.919	.940	.775	.865	.927	.773	.864	.930

Table 27. Correlations between True and Estimated Raw Subscore 3 (GEO)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.924	.961	.979	.763	.865	.924	.761	.864	.926
IM, r=.3	.919	.948	.959	.821	.902	.941	.818	.902	.949
IM, r=.6	.946	.967	.975	.850	.920	.947	.846	.922	.959
IM, r=.9	.962	.977	.983	.871	.928	.949	.867	.931	.966
PM, r=.3	.584	.627	.665	.759	.866	.926	.756	.866	.928
PM, r=.6	.744	.786	.810	.761	.867	.926	.758	.866	.928
PM, r=.9	.880	.919	.940	.757	.865	.926	.754	.864	.929

Table 28. Correlations between True and Estimated Raw Subscore 4 (STA)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.921	.961	.979	.747	.862	.919	.744	.861	.921
IM, r=.3	.922	.949	.961	.821	.904	.940	.818	.904	.947
IM, r=.6	.947	.968	.976	.850	.920	.948	.844	.922	.959
IM, r=.9	.963	.978	.983	.870	.930	.949	.864	.933	.964
PM, r=.3	.565	.614	.637	.754	.860	.920	.750	.859	.922
PM, r=.6	.739	.781	.802	.753	.859	.920	.749	.858	.922
PM, r=.9	.878	.919	.938	.752	.860	.920	.748	.860	.922

Table 29. RMSEs between True and Estimated Raw Total Scores

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	1.854	2.537	3.579	1.799	2.662	3.835	1.942	2.828	3.906
IM, r=.3	1.835	2.483	3.655	1.876	2.714	4.008	2.038	2.890	3.921
IM, r=.6	1.793	2.493	3.921	1.924	2.784	4.337	2.095	2.953	3.975
IM, r=.9	1.761	2.523	4.227	1.957	2.820	4.694	2.131	2.965	4.016
PM, r=.3	1.825	2.581	3.618	1.754	2.604	3.728	1.891	2.746	3.773
PM, r=.6	1.825	2.540	3.539	1.760	2.613	3.741	1.898	2.761	3.804
PM, r=.9	1.836	2.530	3.526	1.779	2.641	3.780	1.920	2.796	3.848

Table 30. RMSEs between True and Estimated Raw Subscore 1 (ARI)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.503	.680	.952	.842	1.253	1.786	.899	1.306	1.781
IM, r=.3	.555	.836	1.393	.891	1.274	1.894	.946	1.321	1.796
IM, r=.6	.521	.759	1.255	.890	1.283	1.990	.945	1.305	1.762
IM, r=.9	.487	.684	1.151	.887	1.276	2.094	.937	1.270	1.760
PM, r=.3	.827	1.581	3.096	.811	1.219	1.762	.866	1.264	1.760
PM, r=.6	.701	1.276	2.428	.814	1.220	1.764	.870	1.266	1.760
PM, r=.9	.556	.871	1.455	.820	1.226	1.764	.878	1.274	1.759

Table 31. RMSEs between True and Estimated Raw Subscore 2 (ALG)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.496	.658	.939	.885	1.299	1.858	.946	1.357	1.863
IM, r=.3	.542	.822	1.422	.890	1.280	1.884	.944	1.330	1.798
IM, r=.6	.496	.744	1.265	.886	1.278	1.982	.936	1.312	1.760
IM, r=.9	.466	.676	1.141	.883	1.290	2.052	.933	1.307	1.739
PM, r=.3	.895	1.664	3.189	.878	1.291	1.826	.937	1.348	1.829
PM, r=.6	.757	1.339	2.509	.877	1.292	1.829	.937	1.350	1.834
PM, r=.9	.570	.874	1.491	.875	1.286	1.825	.935	1.344	1.827

Table 32. RMSEs between True and Estimated Raw Subscore 3 (GEO)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.478	.658	.908	.915	1.323	1.885	.977	1.384	1.901
IM, r=.3	.547	.822	1.432	.908	1.296	1.909	.961	1.345	1.827
IM, r=.6	.500	.744	1.259	.905	1.287	2.006	.957	1.320	1.790
IM, r=.9	.457	.676	1.140	.897	1.287	2.094	.943	1.300	1.747
PM, r=.3	.910	1.664	3.273	.905	1.337	1.888	.965	1.399	1.901
PM, r=.6	.759	1.339	2.577	.901	1.336	1.890	.961	1.399	1.906
PM, r=.9	.558	.874	1.519	.905	1.334	1.884	.967	1.397	1.899

Table 33. RMSEs between True and Estimated Raw Subscore 4 (STA)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.449	.621	.873	.856	1.284	1.881	.912	1.342	1.912
IM, r=.3	.533	.817	1.399	.902	1.278	1.941	.952	1.325	1.868
IM, r=.6	.491	.726	1.234	.908	1.277	2.003	.950	1.307	1.812
IM, r=.9	.451	.658	1.100	.897	1.271	2.092	.936	1.278	1.791
PM, r=.3	.878	1.660	3.223	.868	1.292	1.883	.924	1.351	1.911
PM, r=.6	.729	1.324	2.504	.865	1.297	1.876	.920	1.357	1.909
PM, r=.9	.540	.857	1.468	.870	1.295	1.885	.926	1.353	1.915

Table 34. Reliability of Subscore 2 (ALG)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.848	.923	.958	.600	.753	.861	.597	.752	.866
IM, r=.3	.849	.899	.921	.679	.809	.888	.674	.809	.901
IM, r=.6	.898	.935	.951	.730	.842	.899	.721	.845	.923
IM, r=.9	.925	.955	.966	.762	.859	.902	.752	.866	.932
PM, r=.3	.364	.394	.444	.598	.748	.858	.596	.748	.863
PM, r=.6	.561	.617	.661	.598	.747	.859	.595	.746	.864
PM, r=.9	.775	.844	.884	.600	.748	.860	.597	.747	.865

Table 35. Reliability of Subscore 3 (GEO)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.854	.923	.959	.583	.748	.854	.578	.747	.858
IM, r=.3	.845	.899	.920	.674	.814	.886	.668	.814	.900
IM, r=.6	.895	.935	.950	.722	.846	.897	.716	.850	.921
IM, r=.9	.926	.955	.966	.758	.860	.901	.752	.867	.934
PM, r=.3	.341	.394	.442	.576	.751	.857	.572	.749	.862
PM, r=.6	.554	.617	.655	.579	.752	.857	.574	.751	.861
PM, r=.9	.775	.844	.883	.574	.749	.858	.569	.747	.862

Table 36. Reliability of Subscore 4 (STA)

	CU Calibration			CM Calibration			SU Calibration		
	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
UNI	.849	.924	.959	.559	.743	.844	.553	.742	.848
IM, r=.3	.850	.902	.924	.674	.817	.883	.669	.817	.896
IM, r=.6	.897	.938	.952	.722	.847	.898	.712	.850	.920
IM, r=.9	.927	.956	.967	.758	.865	.900	.747	.870	.930
PM, r=.3	.319	.378	.406	.568	.740	.846	.562	.738	.850
PM, r=.6	.546	.610	.643	.566	.738	.847	.560	.737	.851
PM, r=.9	.771	.845	.879	.566	.740	.846	.559	.739	.849

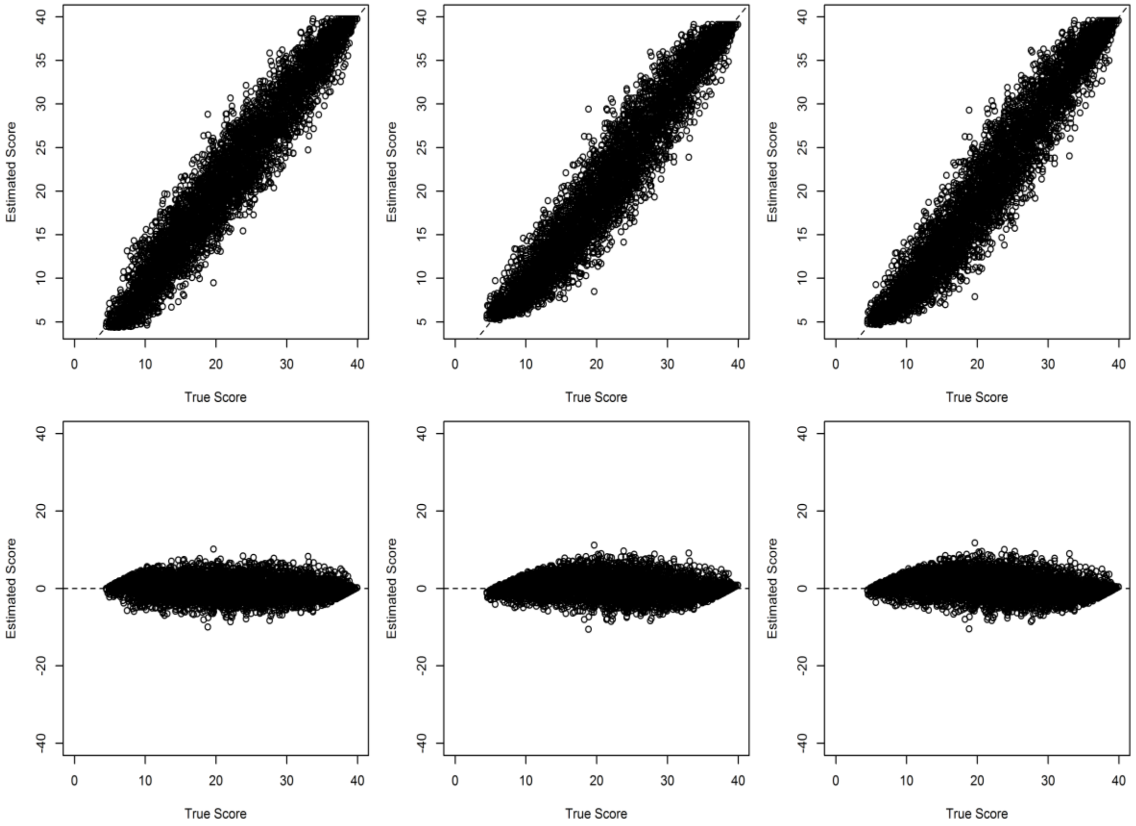

Figure 35. Conditional Biases of Estimated Raw Total Scores for the UNI Design

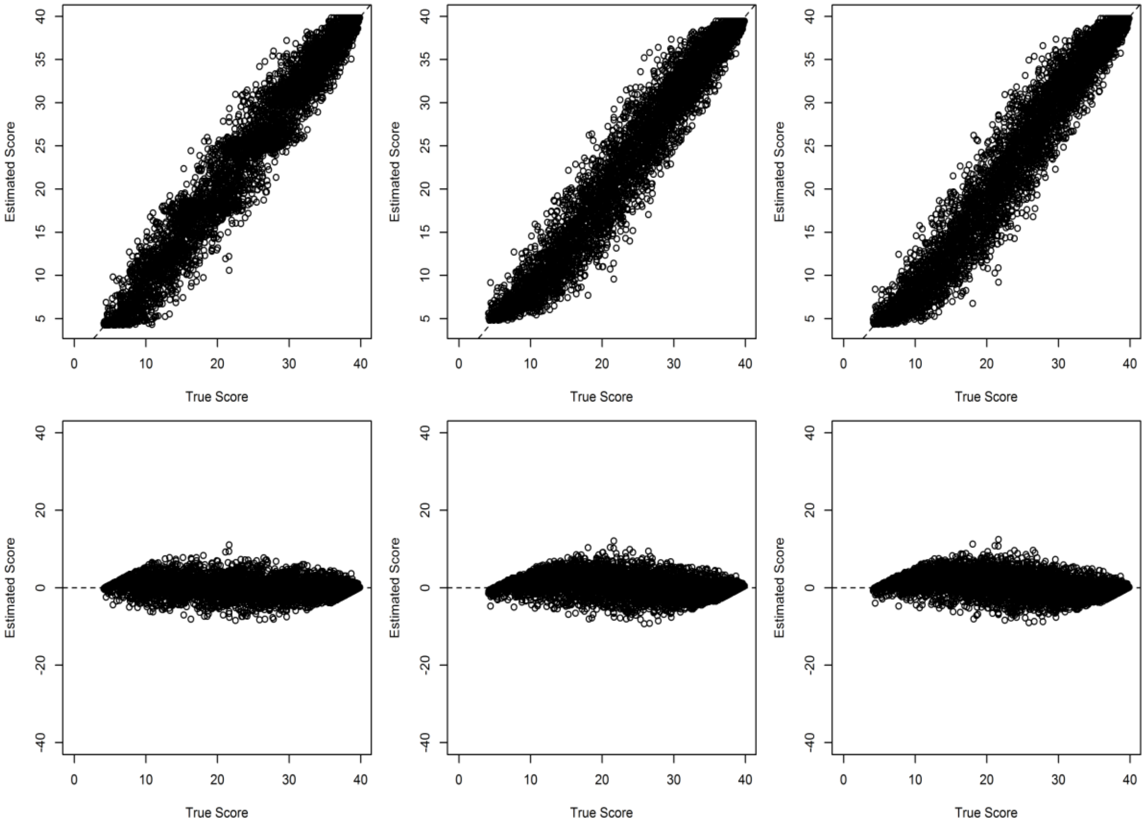

Figure 36. Conditional Biases of Estimated Raw Total Scores for the IM Design with ρ=.3

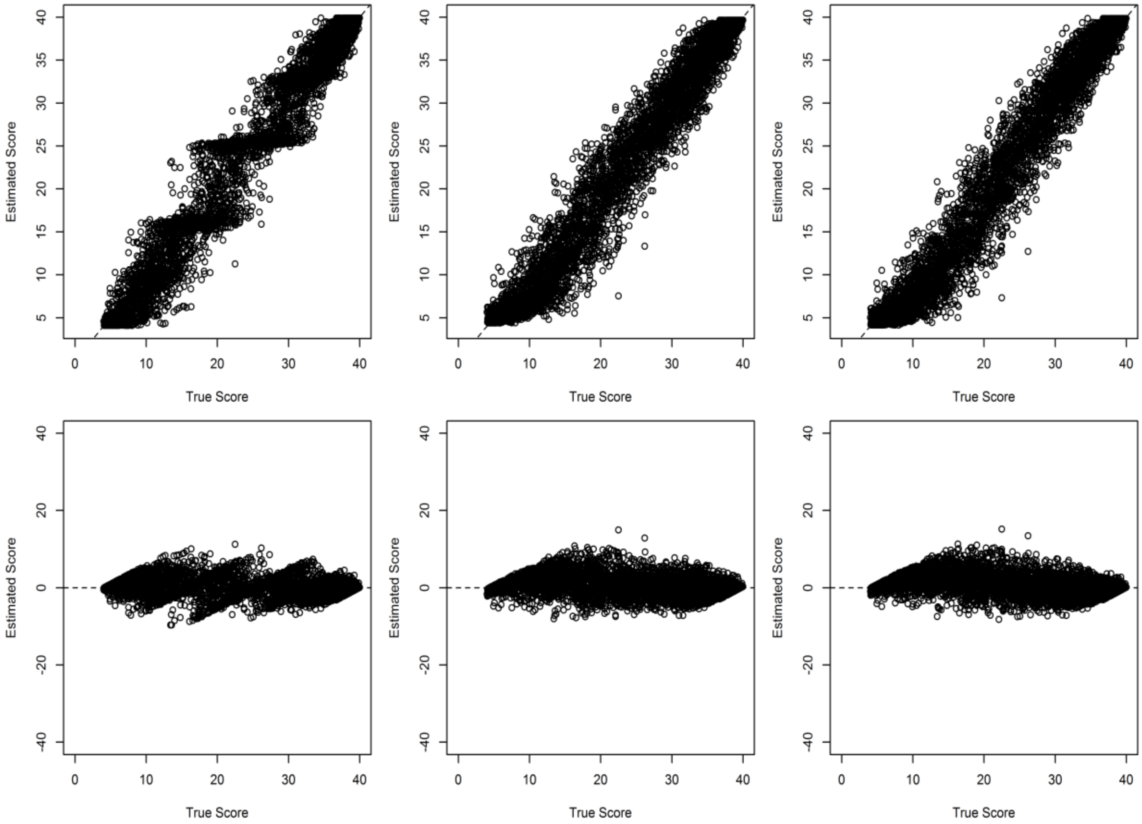

Figure 37. Conditional Biases of Estimated Raw Total Scores for the IM Design with *ρ*=.6

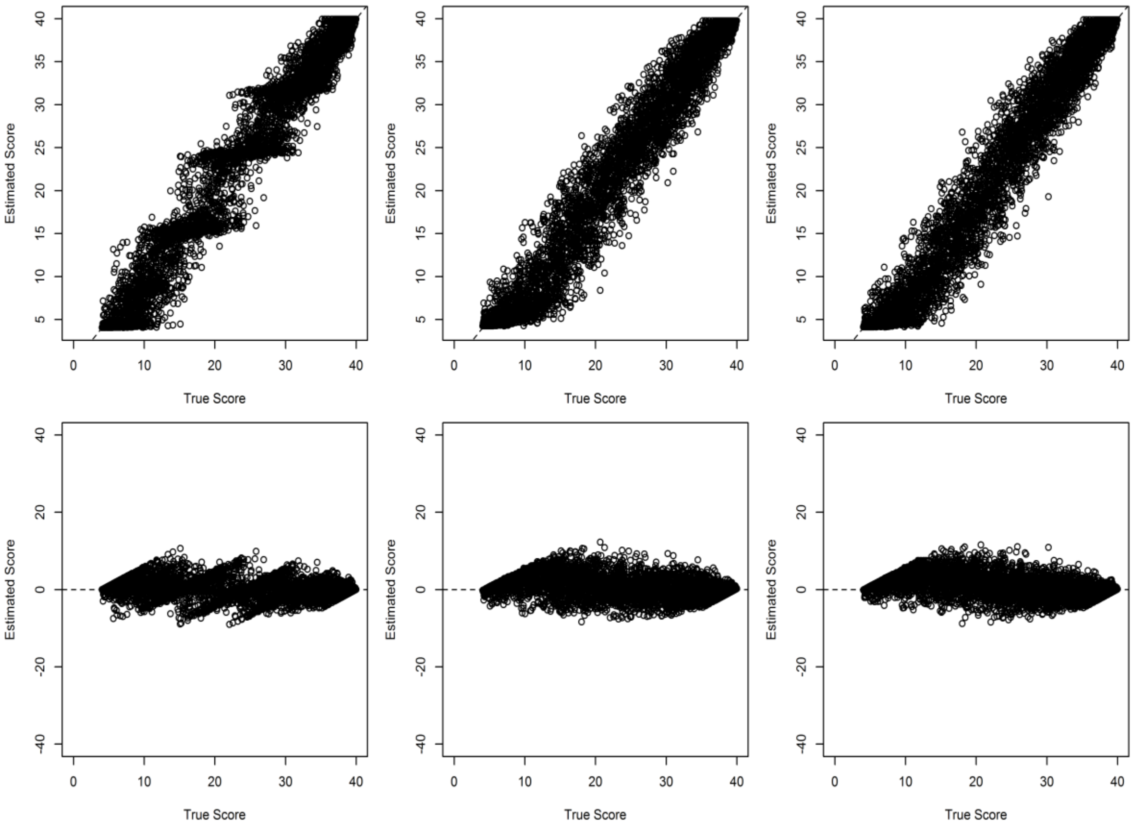

Figure 38. Conditional Biases of Estimated Raw Total Scores for the IM Design with ρ=.9

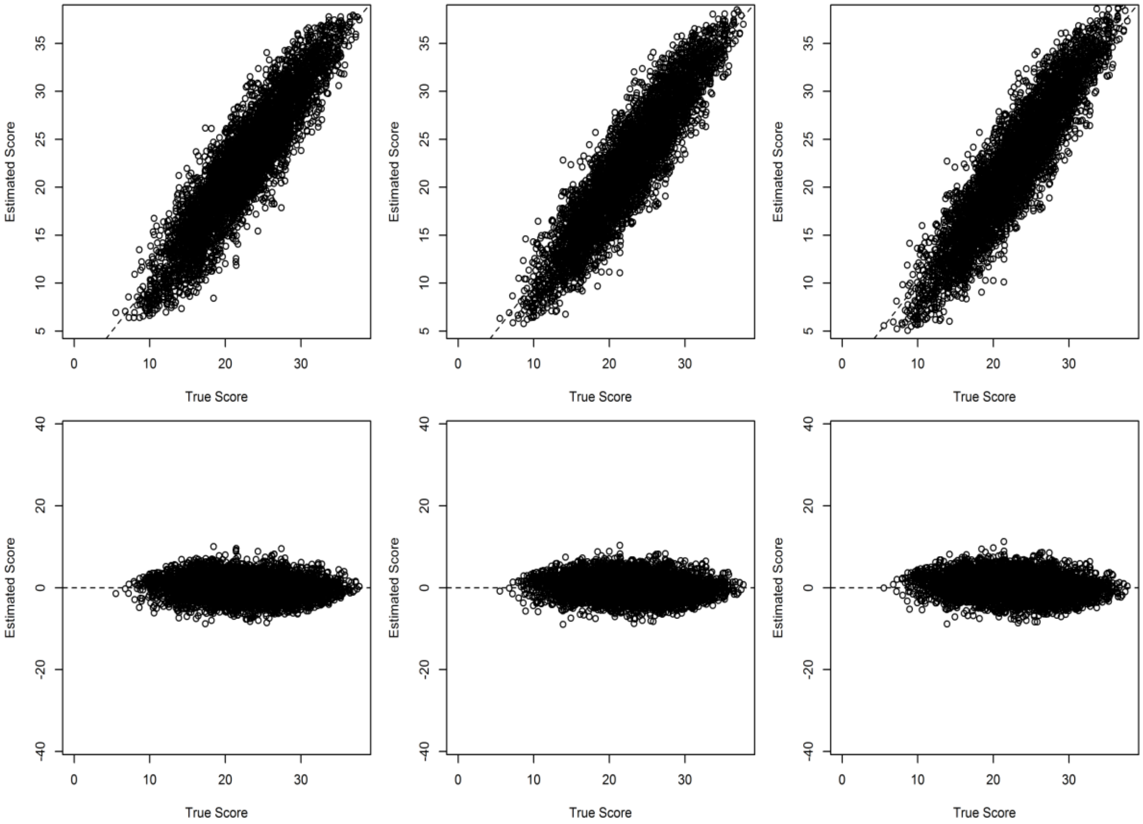

Figure 39. Conditional Biases of Estimated Raw Total Scores for the PM Design with ρ=.3

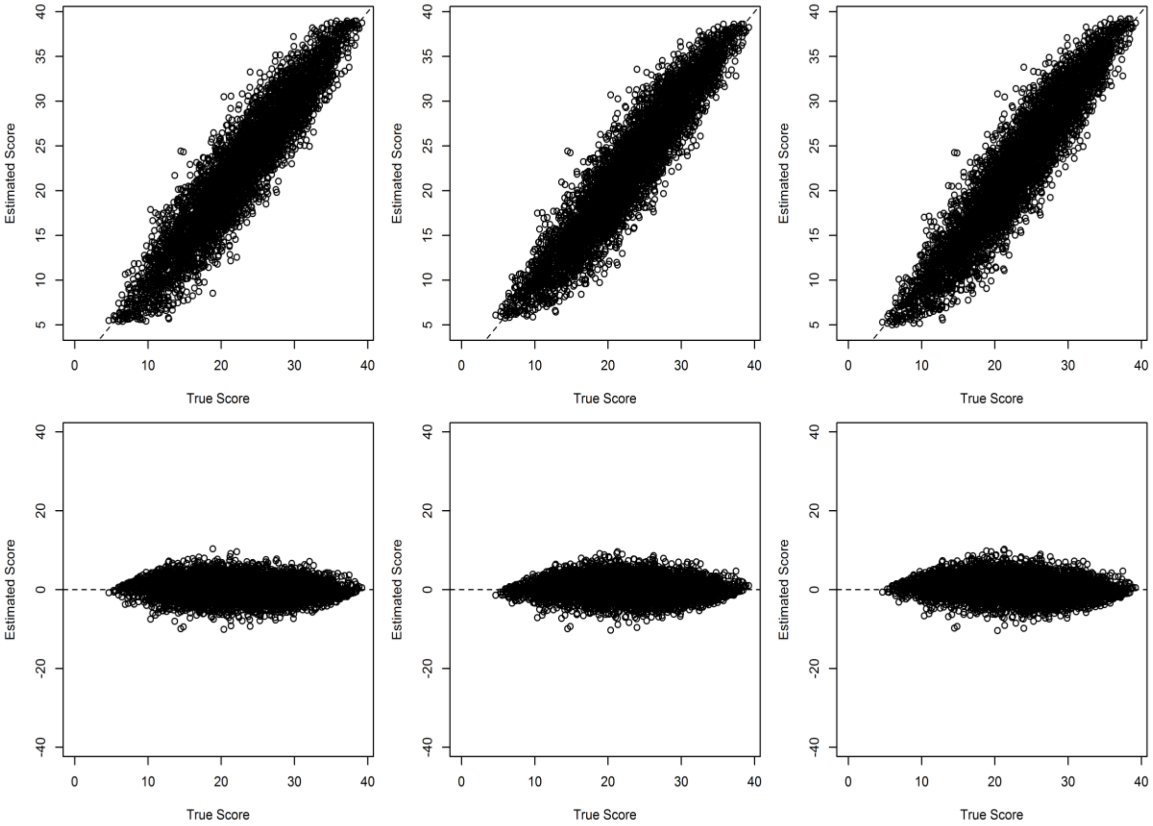

Figure 40. Conditional Biases of Estimated Raw Total Scores for the PM Design with ρ=.6

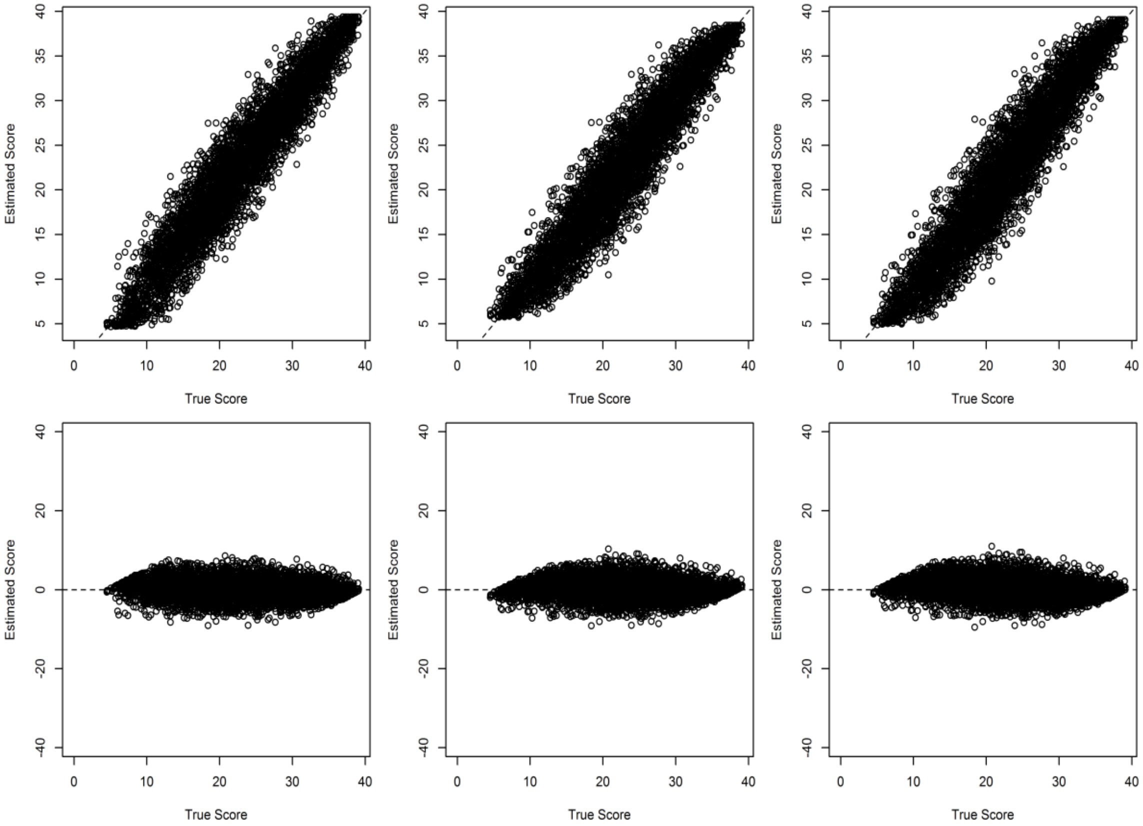

Figure 41. Conditional Biases of Estimated Raw Total Scores for the PM design with ρ=.9

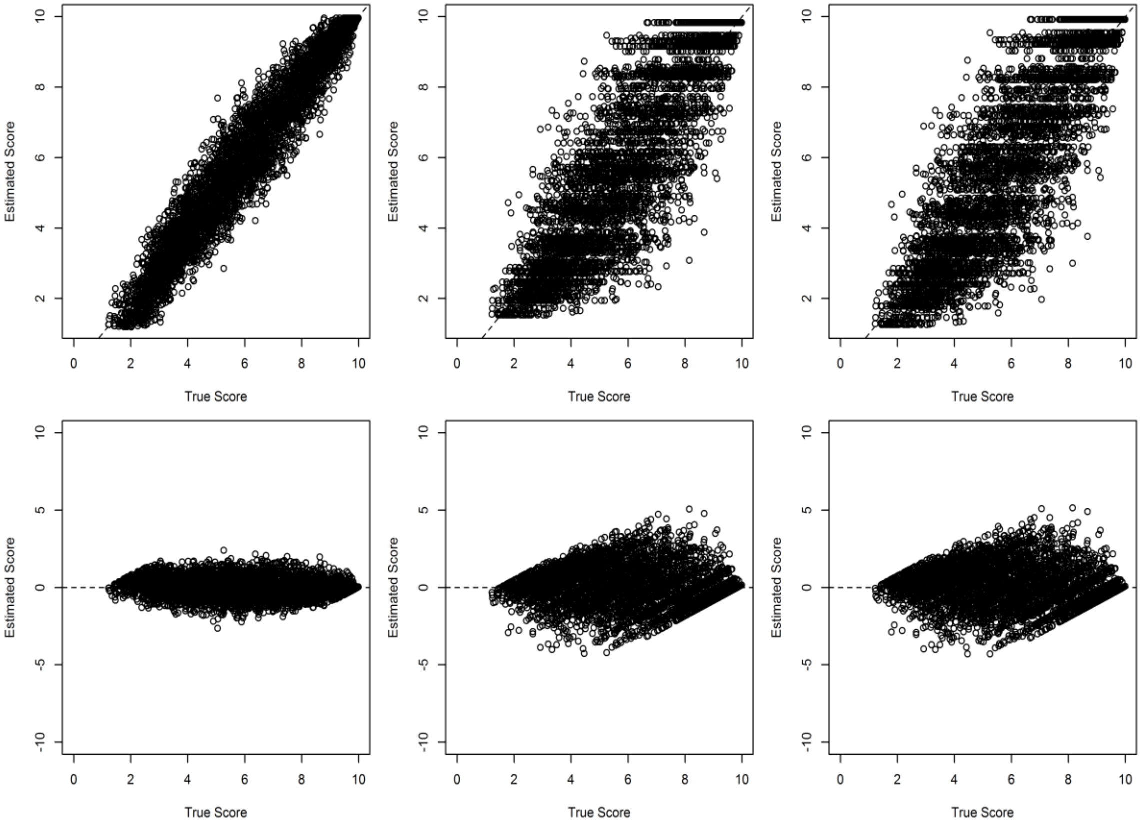

Figure 42. Conditional Biases of Estimated Raw Subscore 1 for the UNI design

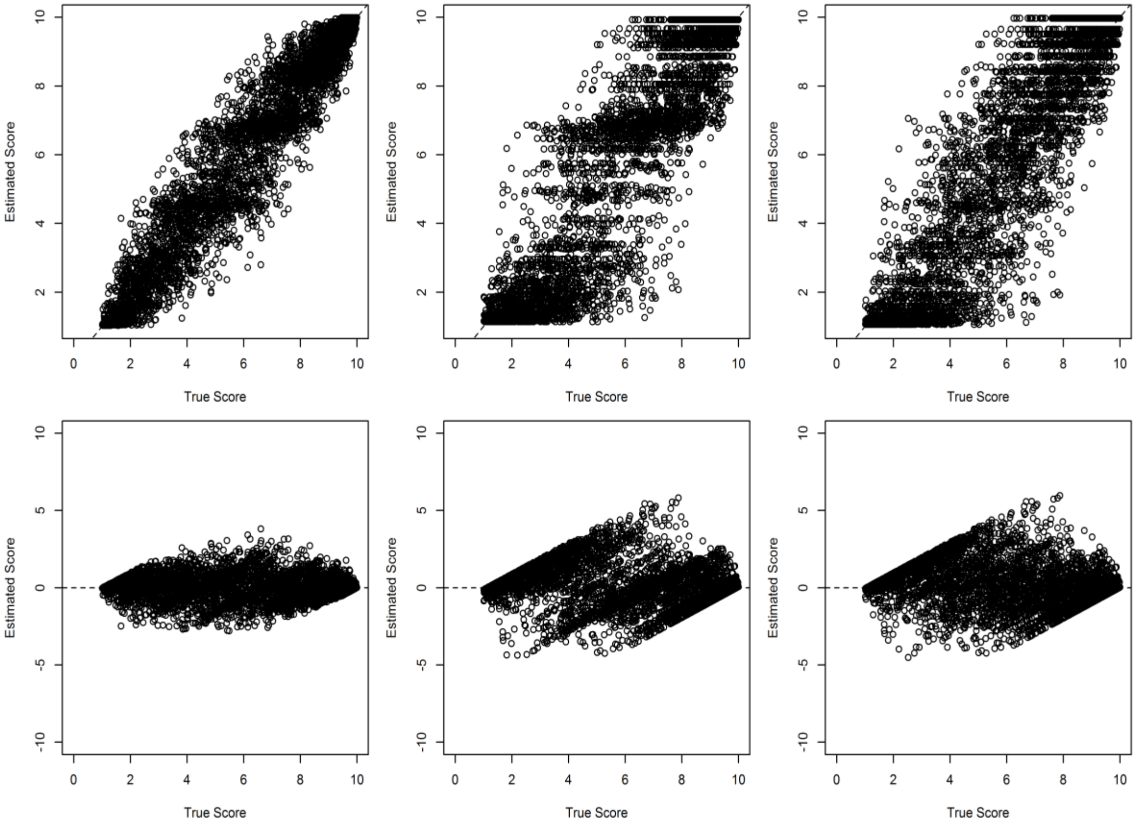

Figure 43. Conditional Biases of Estimated Raw Subscore 1 for the IM Design with *ρ*=.3

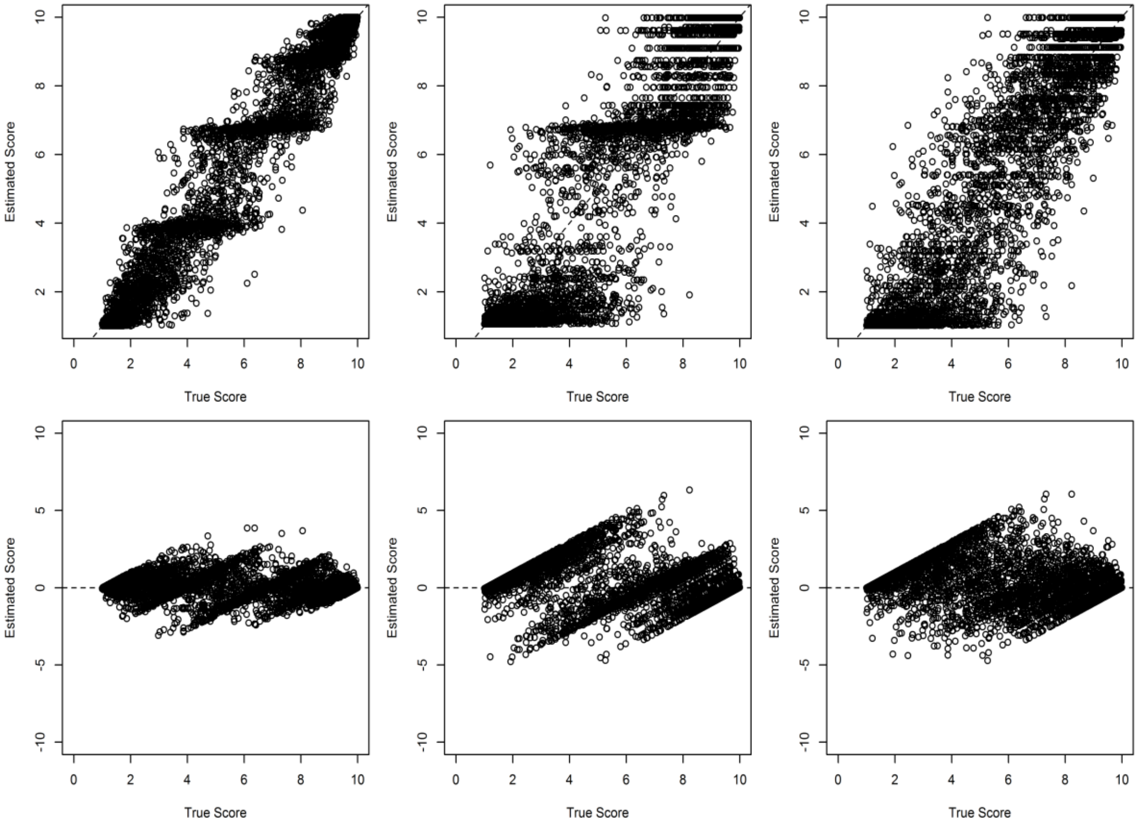

Figure 44. Conditional Biases of Estimated Raw Subscore 1 for the IM Design with ρ=.6

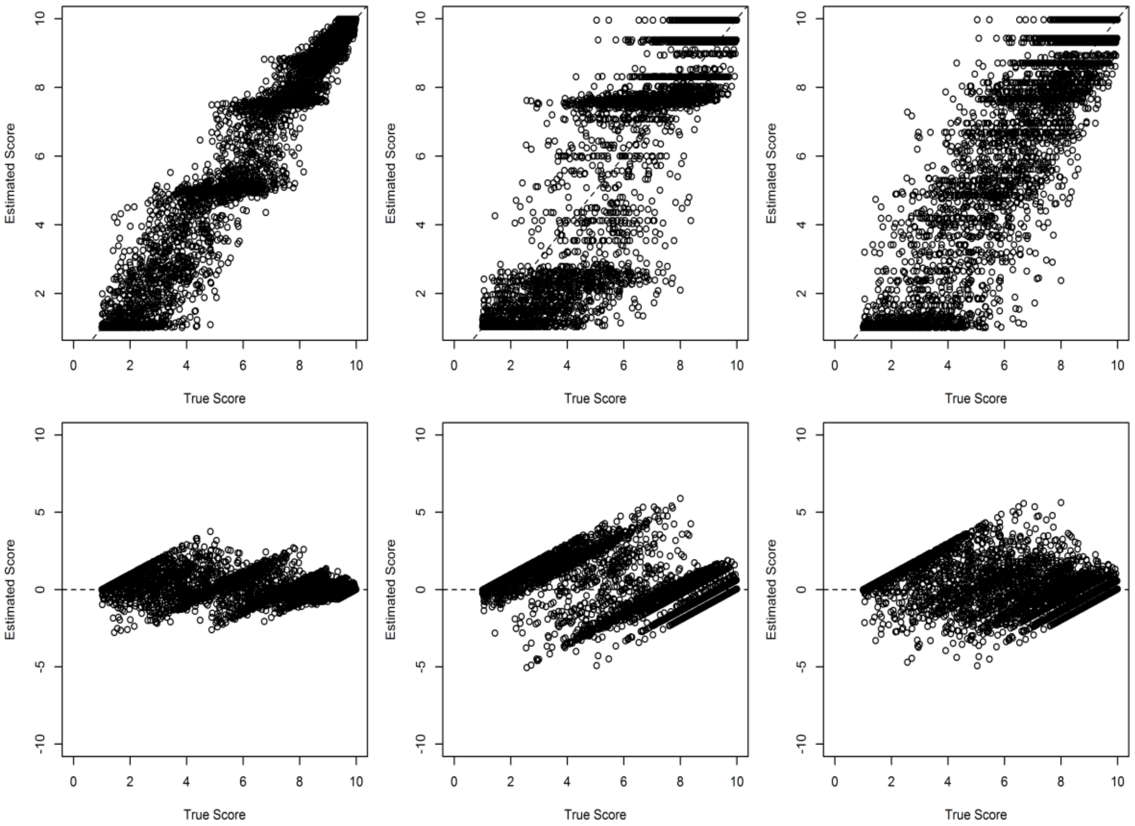

Figure 45. Conditional Biases of Estimated Raw Subscore 1 for the IM Design with *ρ*=.9

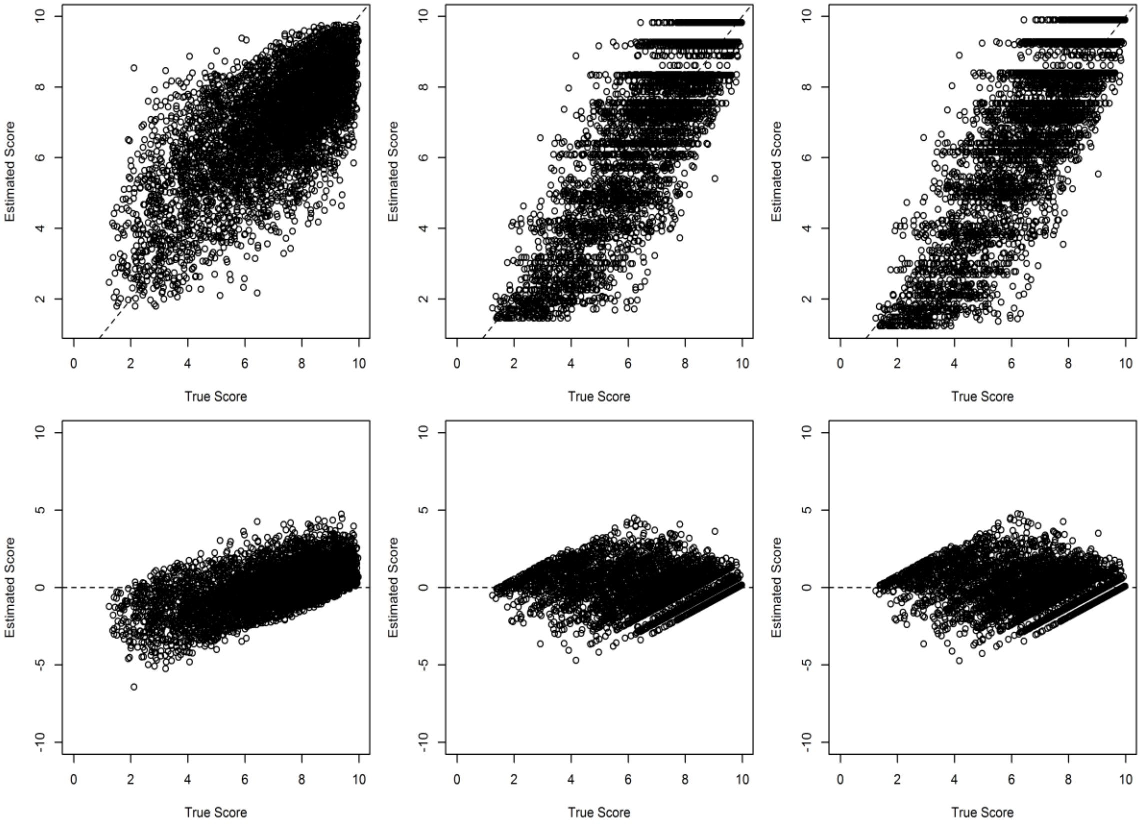

Figure 46. Conditional Biases of Estimated Raw Subscore 1 for the PM Design with ρ=.3

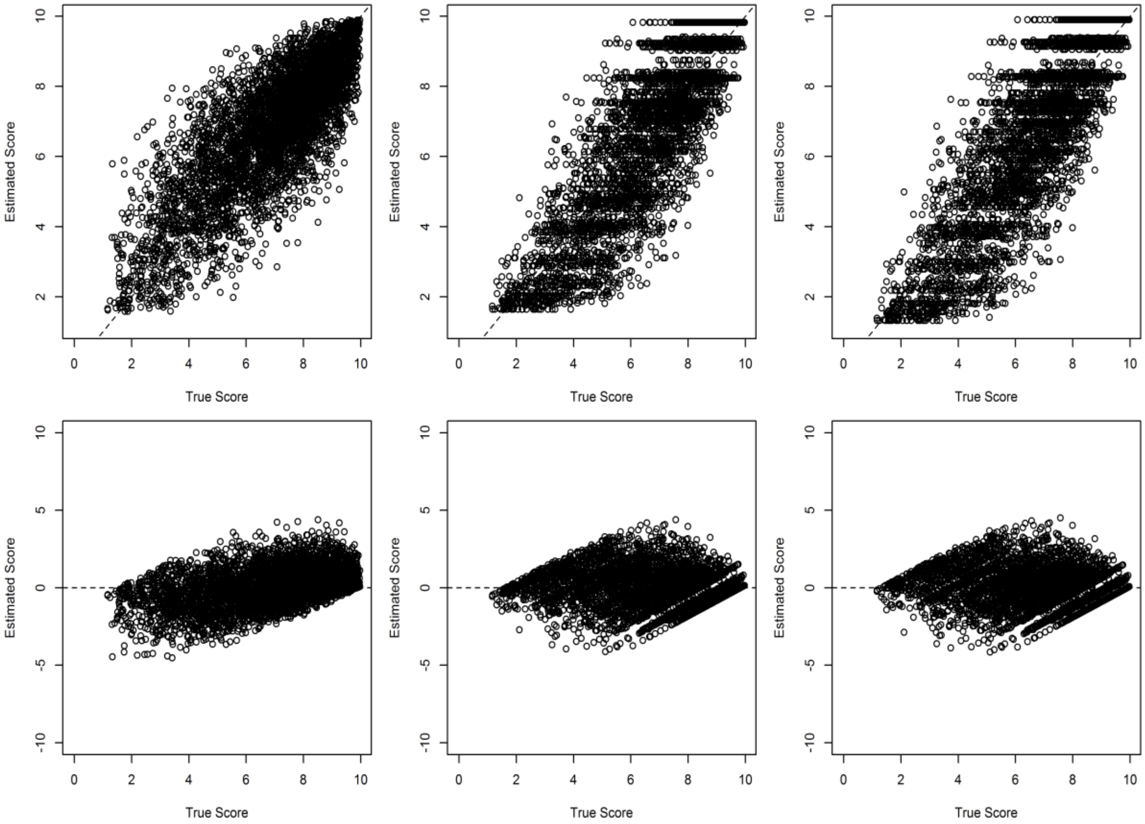

Figure 47. Conditional Biases of Estimated Raw Subscore 1 for the PM Design with ρ=.6

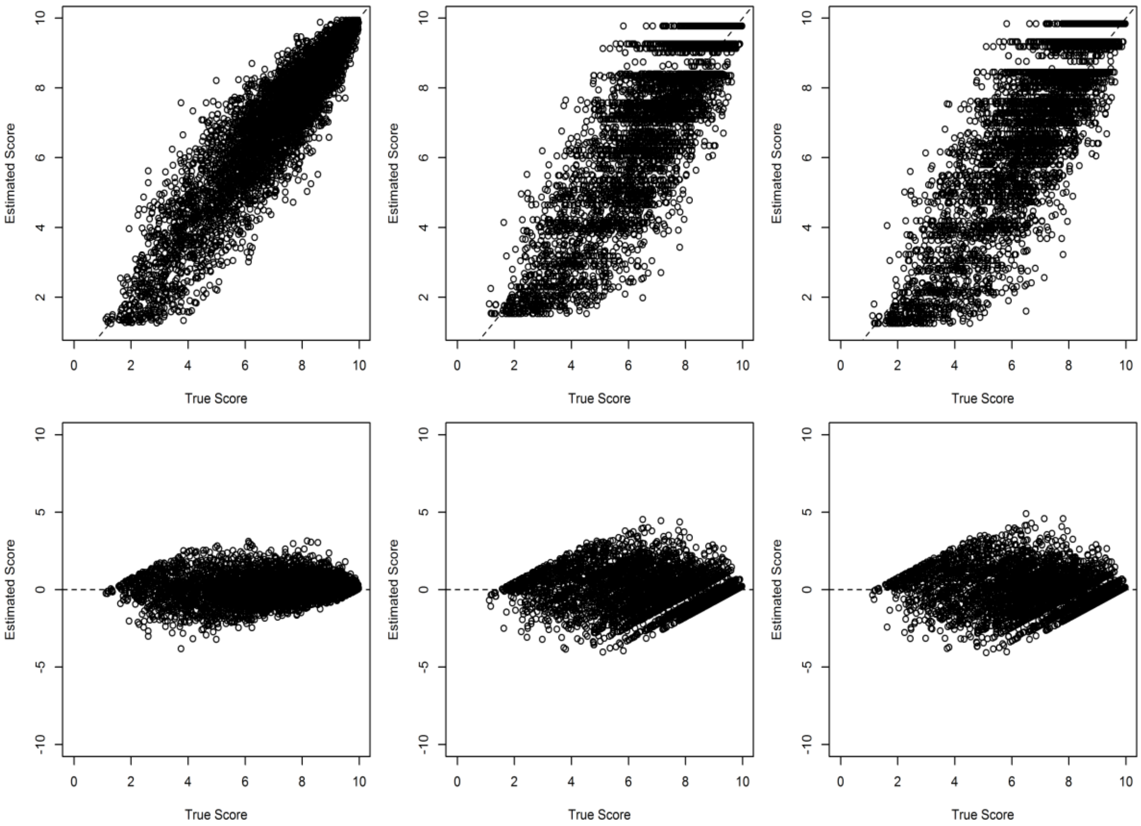

Figure 48. Conditional Biases of Estimated Raw Subscore 1 for the PM Design with ρ=.9

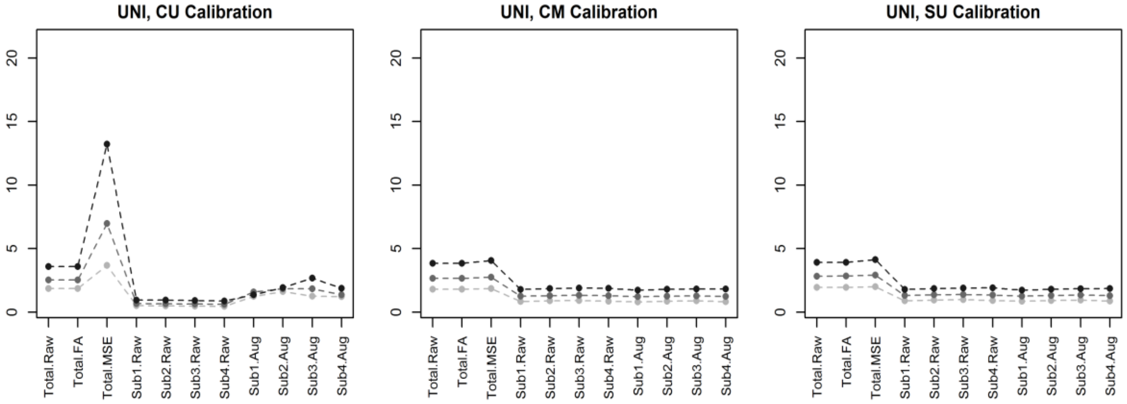

Figure 49. RMSEs between True and Estimated Scores for the UNI Design

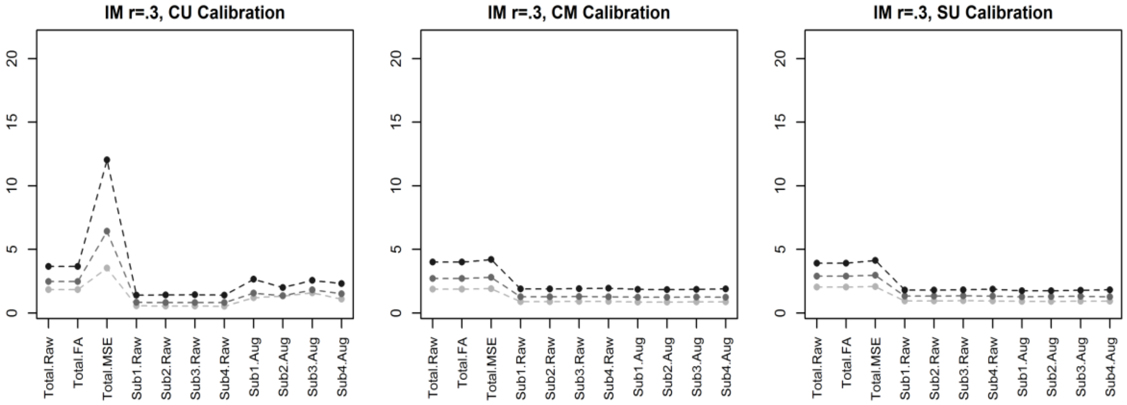

Figure 50. RMSEs between True and Estimated Scores for the IM Design with ρ=.3

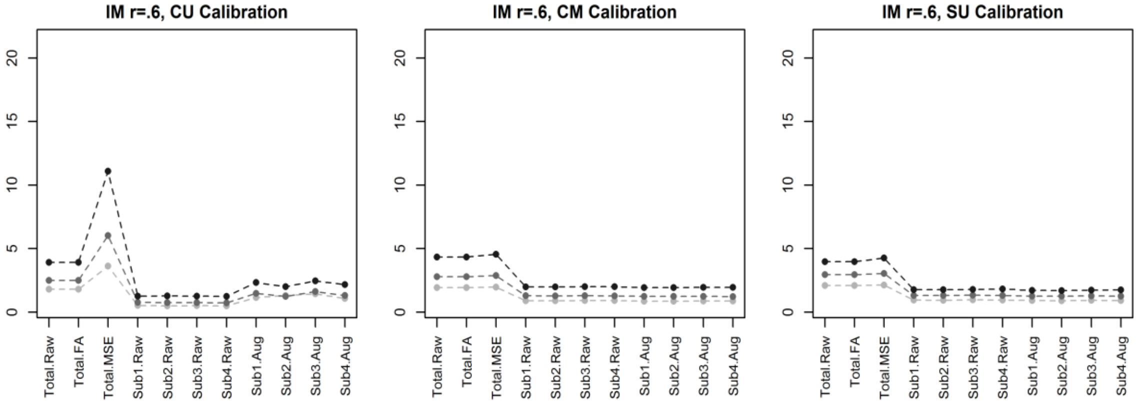

Figure 51. RMSEs between True and Estimated Scores for the IM Design with *ρ*=.6

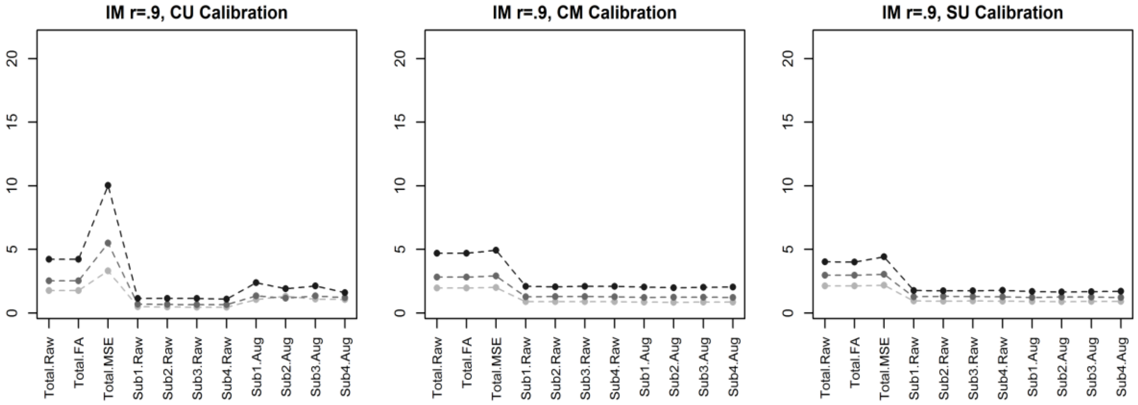

Figure 52. RMSEs between True and Estimated Scores for the IM Design with *ρ*=.9

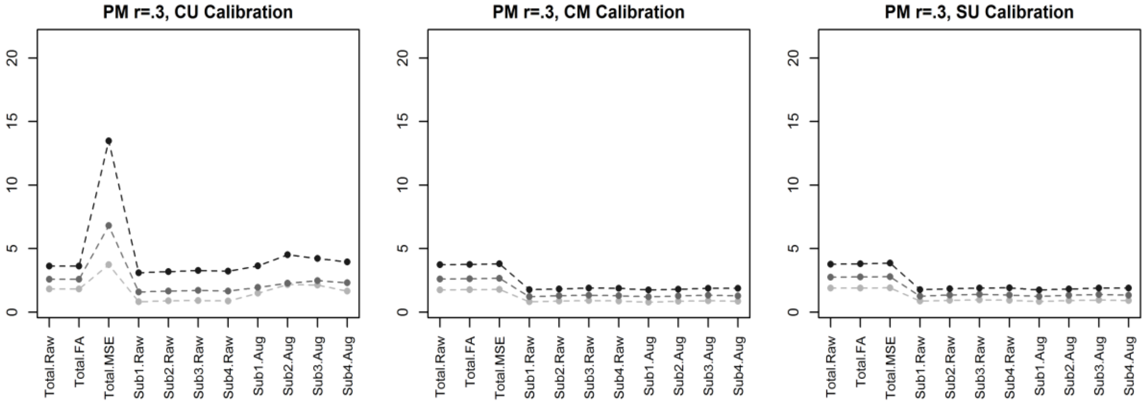

Figure 53. RMSEs between True and Estimated Scores for the PM Design with *ρ*=.3

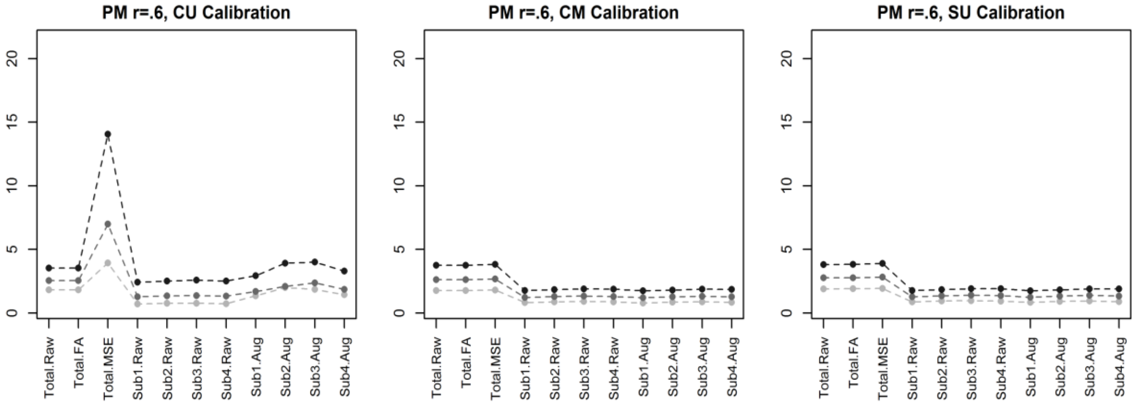

Figure 54. RMSEs between True and Estimated Scores for the PM Design with *ρ*=.6

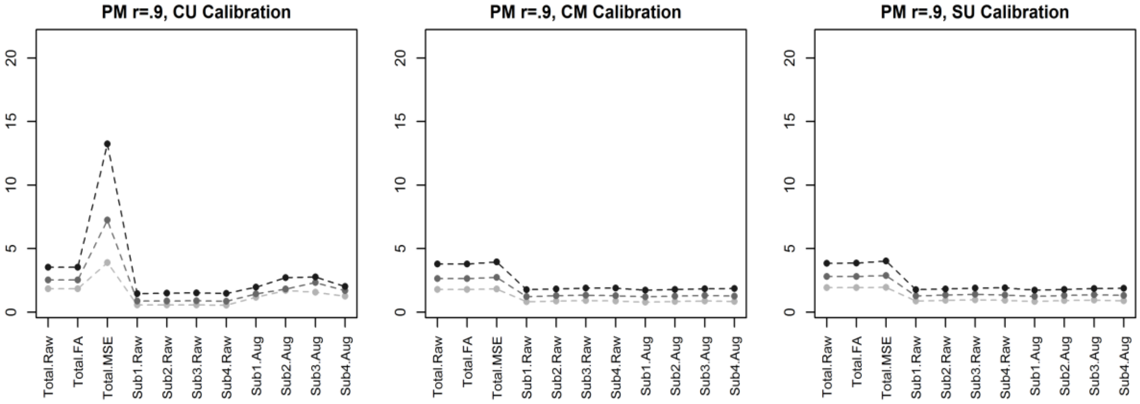

Figure 55. RMSEs between True and Estimated Scores for the PM Design with *ρ*=.9