LIU, FU, Ph.D. Comparisons of Subscoring Methods in Computerized Adaptive Testing: A Simulation Study. (2015)
Directed by Dr. Terry A. Ackerman. 187 pp.

Given the increasing demands of subscore reports, various subscoring methods and augmentation techniques have been developed aiming to improve the subscore estimates, but few studies have been conducted to systematically compare these methods under the framework of computerized adaptive tests (CAT). This research conducts a simulation study, for the purpose of comparing five subscoring methods on score estimation under variable simulated CAT conditions. Among the five subscoring methods, the IND-UCAT scoring ignores the correlations among subtests, whereas the other four correlation-based scoring methods (SEQ-CAT, PC-MCAT, reSEQ-CAT, and AUG-CAT) capitalize on the correlation information in the scoring procedure. By manipulating the sublengths, the correlation structures, and the item selection algorithms, more comparable, pragmatic, and systematic testing scenarios are created for comparison purposes. Also, to make the best of the sources underlying the assessments, the study proposes a successive scoring procedure according to the structure of the higher-order IRT model, in which the test total score of individual examinees can be calculated after the subscore estimation procedure is conducted. Through the successive scoring procedure, the subscores and the total score of an examinee can be sequentially derived from one test.

The results of the study indicate that in the low correlation structure, the original IND-CAT is suggested for subscore estimation considering the ease of implementation in practice, while the suggested total score estimation procedure is not recommended given

the large divergences from the true total scores. For the mixed correlation structure with two moderate correlations and one strong correlation, the original SEQ-CAT or the combination of the SEQ-CAT item selection and the PC-MCAT scoring should be considered not only for subscore estimation but also for total score estimation. If the post-hoc estimation procedure is allowed, the original SEQ-CAT and the reSEQ-CAT scoring could be jointly conducted for the best score estimates. In the high correlation structure, the original PC-MCAT and the combination of the PC-MCAT scoring and the SEQ-CAT item selection are suggested for both the subscore estimation and the total score estimation. In terms of the post-hoc score estimation, the reSEQ-CAT scoring in conjunction with the original SEQ-CAT is strongly recommended. If the complexity of the implementation is an issue in practice, the reSEQ-CAT scoring jointly conducted with the original IND-UCAT could be considered for reasonable score estimates.

Additionally, to compensate for the constrained use of item pools in PC-MCAT, the PC-MCAT with adaptively sequencing subtests (SEQ-MCAT) is proposed for future investigations. The simplifications of item and/or subtest selection criteria in a simple-structure MCAT, PC-MCAT, and SEQ-MCAT are also pointed out for the convenience of their applications in practice. Last, the limitations of the study are discussed and the directions for future studies are also provided.

COMPARISONS OF SUBSCORING METHODS IN COMPUTERIZED ADAPTIVE TESTING: A SIMULATION STUDY

by

Fu Liu

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
2015

Approved by

____
Committee Chair

DEDICATION

To everyone who encourages, inspires, and truly loves me, particularly my mother, Hongyan Zhang.

APPROVAL PAGE

This dissertation, written by Fu Liu, 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

ACKNOWLEDGEMENTS

An internship project triggered my strong interests in computerized adaptive tests (CAT). It was so intriguing that I was hoping that I could expand it in the dissertation. However, this choice might not be wise in most of eyes. One reason was that I had been in the dissertation focusing on another topic (MIRT) for a while. I was more familiar with that topic and also had already done lots of work on it. Conversely, CAT was quiet a new arena to me. The work I had done relevant to CAT was only that internship project. In the meantime, the people around tried to convince me that we have always to make some more practical choices in our lives. I could not agree with them more, and then I was asking myself, "Can I make a few decisions only for my own in this life?" For the dissertation, I wanted to follow my own heart, although I knew I would have to pay more for it, maybe at the moment or maybe in the future.

Thanks to my advisor, Dr. Terry Ackerman, his supports and encouragements strengthened my determination for the change. I would also like to thank him for his advice on my dissertation and his persistent efforts of organizing and coordinating the committee members. My sincere thanks go to Dr. Wim van der Linden who always dragged me back when my work was out of track, and then provided me new clues of thinking that improved the quality of the dissertation. I also want to thank my other committee members, Dr. Robert Henson, Dr. Richard Luecht, and Dr. Sat Gupta, for their careful review and valuable comments, so that the dissertation was continuously polished.

I would also like to expand my appreciation to Dr. Haimeng Zhang for his interesting and enlightening discussions with me on math and statistics.

In one word, what I can express here is far less than enough compared to my deep and sincere gratitude to all the professors for their patience, generosity, and tolerance. In the meantime, the more I explore on this topic, the more interesting things I have found and will find on CAT. However, for the dissertation, I have to stop somewhere anyway, so this is it.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Background

Test scores are typically perceived as the informative evidence of reflecting the performance of examinees in a test and also as an important reference of predicting examinees' future academic or career growth trajectory. Accordingly, the quality of the reported scores, which is mainly evaluated by the fairness, validity, and reliability of the scores, becomes a critical concern as the scoring procedure is implemented. These properties of test scores remarkably determine the significance and accountability of educational assessments. Nowadays, the rapid development of scoring techniques largely guarantees fair, valid and reliable total scores for the purpose of making high-stake decisions, such as college admission, promotion screening, and professional licensure. The total score of a test reflects summative assessment, which aims to evaluate the examinees' overall performances in an entire test and differentiate their placements and proficiency levels on the general latent trait scale continuum. In the recent years, there has been a rising demand for subscores in the testing market. Subscores manifest formative/interim assessment, which seeks to monitor the performances of examinees at the level of some specific subscales or content areas and derive the diagnostic feedbacks on them for future learning and teaching modifications.

In practice, most of modern large-scale assessments are intended to measure a general latent trait or a broad subject, which are always broken down to some particular content areas, instructional objectives, or subscale categories in a curriculum and test design. For instance, a state science achievement test may consist of four subtests covering four content areas, which are Nature of Science, Biological Sciences, Physical Sciences, and Earth and Space Sciences. This type of test construction structure is very common in educational and psychological assessments and is basically recognized as the hierarchical latent trait structure in the modern test theory. For such tests, the most widely used operational approach of deriving test scores is to apply a unidimensional item response theory (UIRT) model to estimate IRT general ability parameters and then convert them to interpretable scale scores, which are ultimately reported as total scores to the public. However, to ensure that students meet the standards of state assessments, teachers, students, parents and even school administration officers gradually show more concerns on subscore reports, which provide diagnostic information regarding different content areas or instructional objectives, in order to be aided in locating the strengths and weaknesses of students for future instructional and learning remediation.

Subscores are also known as domain scores, diagnostic scores, subscale scores, and objective-level scores (e.g., de la Torre & Song, 2009; Sinharay, Puhan, & Haberman, 2010; Stone, Ye, Zhu, & Lane, 2010; Skorupski & Carvajal, 2010). For examinees, particularly failing candidates, subscores explicitly reflect their strengths and weaknesses and are of great benefit to them to accordingly adjust their future study directions. Subscores could also assist classroom teachers to plan individual remedial instructions and to track down gaps between teaching and learning among different instructional objectives. Based on subscore reports, state educational institutions could evaluate the quality of their curriculum and the effectiveness of teaching and learning in a finer-grained manner. Other than giving diagnostic information, subscores could also provide additional information in conjunction with total scores to some interested parties (e.g., admission or funding officials and company employers), allowing them to screen all qualified candidates for the one(s) with some unique strong skill(s) that can specifically complement their team (Monaghan, 2006). Therefore, the usefulness of subscores is apparent and non-negligible for different layers of interested parties.

In the National Research Council report “Knowing What Student Know” (2001), it was stated that “To do justice to the students in our schools and to support their learning, we need to recognize that the process of appraising them fairly and effectively requires multiple measures constructed to high standards. Useful and meaningful evidence includes profiling of multiple elements of proficiency, with less emphasis on overall aggregate scores” (p. 313). The report also encouraged assessment developers to fully exploit the advanced technology “to assess what students are learning at fine levels of detail, with appropriate frequency, and in ways that are tightly integrated with instruction” (p. 306). The No Child Left Behind (NCLB) Act of 2001 (U.S. Department of Education, 2002) addressed that the state academic assessments required to “produce individual student interpretive, descriptive, and diagnostic reports” (§1111. p. 1451) for teachers, parents, and principals to better specify academic needs of students. Such reports, currently circulated in different states, were refined by Goodman and Hambleton (2004) as two primary categories. One is to present the assessment outcomes (e.g. raw scores or percentile rank scores) in terms of the students' attainable knowledge or skills on some subdomains. The other is to enumerate the specific knowledge or cognitive skills required to be improved in the future. The subscores investigated in this study belong to the former.

In 2010, the release of the Common Core standards for mathematics and English language arts (the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO), 2010) marked a new start of standards-based education reform in the United States. Across diverse state curricula, the standards specify and describe the skills and knowledge that students are expected to acquire within the subjects of mathematics and English language art at each grade level. Inspired by the Common Core Standards, subscore reports on specific skills or knowledge are anticipated to be more highly desirable in the near future. There is no doubt that as subscore reports are increasingly demanded as an important assessment outcome, attentions to the quality of subscores must be growing. Currently, some large-scale testing programs such as ACT, LSAT, and SAT provide subscore reports to examinees. However, in the face of the present testing circumstances, the development and extensive applications of subscoring are still very restricted due to some potential challenges.

In policy, Standard 5.12 in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) clearly illustrates that "Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established" (p.65). The National Research Council report "Knowing What Student Know" (2001) also emphasizes that "Assessment designers should explore ways of using sets of tasks that work in combination to diagnose student understanding while at the same time maintaining high standards of reliability" (p.306).

In practice, some crucial studies addressed the potential crisis that some reported subscores, such as raw scores, proportion correct scores, or percentile rank scores (Crocker & Algina, 1986), might be inaccurate, unreliable, and even somewhat meaningless, especially when the original test specifications were not designed for subscoring (e.g., Goodman & Hambleton, 2004; Haberman, 2008; Dorans, 2005; Tate, 2004; Sinharay & Haberman, 2008). If such subscores are reported, they may mislead test users and audiences and result in misinterpretation of examinees' performance in subtests. Specifically speaking, a test evaluating examinees on a general ability or subject may consist of a test battery that includes several subtests measuring different subscales or content areas. The intended use of such assessments is typically to rank examinees on the general ability scale instead of on the subscales. One of the principles considered when these tests are designed is to ensure valid and reliable total scores on the basis of cost efficiency. The considerations on the cost and time invested result in a dilemma that the items used in each subtest for measuring specific content areas or subscales are very limited. As a consequence, subscores, when estimated by traditional scoring methods, are not adequately reliable and accurate and must be reported and interpreted with caution.

Confronted with these challenges, a large number of studies focused their attention on improving and developing subscore estimation approaches. Under the framework of classical test theory (CTT), some researchers predicted the true subscore by regressing it on the observed subscores or observed total scores or both (Wainer et al., 2001; Haberman, 2008). Wainer et al. (2000, 2001) applied a similar regression approach on different types of IRT scale scores and developed the augmented subscores (AUG) through borrowing information from the other subtests. Another empirical Bayes (EB) subscoring procedure is known as Objective Performance Index (OPI; see Yen, 1987), which combines the informative prior ability distribution obtained from the entire test to the observed subscore estimates.

Furthermore, given the fact that multiple subscales are measured in one test, some researchers embedded the subscoring procedure to the multidimensional IRT (MIRT) models by using Markov Chain Monte Carlo (MCMC) estimation techniques (de la Torre & Patz, 2005; Sheng & Wikle, 2007). Based on the nature of the hierarchical latent trait structure in the test, the higher-order IRT (HO-IRT) model was developed, which can simultaneously estimate total scores and subscores (de la Torre & Song, 2009; Huang, Wang, Chen, & Su, 2013). There were some other studies considering ancillary information such as demographic factors to improve the accuracy of IRT subscore estimates (de la Torre, 2009). In addition, a few studies examined the efficiency of some subscoring methods on polytomous item responses (de la Torre, 2008; Yen, Sykes, Ito, & Julian, 1997; Shin, 2007; Wang & Chen, 2004).

Looking over all the methods mentioned above, the core concept mainly focuses on how to make full use of the information collateral to the other subtests, so as to realize the improvement of the subscore estimation in the target subtest. The rationale of doing this rests on the fact that the collateral information could, to some extent, compensate for short subtest length and improve the accuracy and reliability of subscores. Substantial studies have verified the argument and asserted that these methods outperform the traditional classical and IRT subscoring approaches that do not utilize collateral information, such as the proportion-correct (PC) observed subscores and multiple independent UIRT subscore estimation (Edwards & Vevea, 2006; Wainer et al., 2001; Kahraman & Kamata, 2004; Yen, 1987; Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; DeMars, 2005).

Motivation of the Study

To further examine the performance of these methods, numerous studies were implemented aiming to evaluate their relative performances as some relevant influential factors alter. Table 1 presents some subscoring approaches that were primarily included in the comparison for the paper-and-pencil (P&P) tests in the literature. It is apparent that Wainer's AUG (which refers to Wainer et. al's AUG for brevity in this study), Yen's OPI, and MIRT estimation are the most widely acknowledged subscoring methods in P&P tests, while the PC subscoring is usually treated as the representative of traditional estimation approaches and the baseline method. The most recent study (de la Torre, Song, & Hong, 2011) also added the higher-order IRT model into the comparison, based on the fact that the hierarchical latent trait structure is often adopted in the design of modern assessments. In the meantime, some crucial factors that may affect subscore estimation were investigated in most of the previous simulation studies, which are listed in Table 2. Typically, they refer to the number of content areas or instructional objectives, subtest

Table 1

Primary Comparison Studies on Subscore Estimation Approaches for P&P Tests in the Literature

		Methods for Comparison			
Author	Model for Data Generation or Item Calibration	Wainer et. al's AUG	Yen's OPI	MIRT	Other Methods
DeMars (2005)	3PL MIRT model	√		√	Bi-factor model and independent UIRT models.
Dwyer et al.(2006)[a]	UIRT	√	√	√	PC subscores.
Edwards & Vevea (2006)	3PL UIRT model	√ (AUG on observed summed scores and IRT scale scores for summed scores)			
Shin (2007)	3PL UIRT model and the generalized partial credit model	√	√		The Bock's method (1997), the Shin's method (2005), and PC subscores.
Yao & Boughton (2007)	3PL MIRT model and multidimensional partial credit model		√	√	PC subscores, UIRTOJSS, MIRTPSS, BMIRTSS, BMIRTDS (2007)
Stone et al. (2009)[a]	Not mentioned	√	√	√	
Fu & Qu (2010)	Multidimensional partial credit model	√	Adjusted OPI	√	PC observed subscores, PCM_SUB, PCM_ALL (2010)
Skorupski & Carvajal (2010)	3PL UIRT model	√ (AUG on CTT raw scores & AUG on IRT raw scores)			A Bayesian IRT with Informative Priors Approach.
de la Torre et al. (2011)[a]	Higher-order IRT model	√	√	√	Higher-order IRT models.

Note. [a]For the sake of space in the table, only the first author in their study is listed. Please check the reference list for detailed information.

Table 2

Primary Factors Affecting Subscore Estimation Investigated for P&P Tests in the Literature

Author	number of subtests	Test length	Subscale correlations	Sample size	Simulation	number of replication	Real data
DeMars (2005)	2	20 for one subtest and 15 for the other	0.81	2,552 in real data and 2,500 for simulation	√	100	√
Dwyer et al. (2006)[a]	Uni-data: Reading: 4; Mathematics: 7 Multi-data: 5	Uni-data: Reading: 32 altogether for 4 subtests; Mathematics: 31 altogether for 7 subtests. Multi-data: 5 & 10 for each subtest		Uni-data: Reading: 1,983 Mathematics: 1,430 Multi-data: 6,000			√
Edwards & Vevea (2006)	2 & 4	Different combinations of 5, 10, 20, & 40 items for each subtest (See p. 245)	0.3, 0.6, & 0.9	2,000	√	100	
Shin (2007)[b]	Not mentioned	6, 12, & 18 for each subtest	0.5, 0.8, & 1.0	250, 500, & 1,000	√	100	
Yao & Boughton (2007)	4	60 dichotomous and polytomous items altogether for 4 subtests	0, 0.1, 0.3, 0.5, 0.7, & 0.9	1,000, 3,000, & 6,000	√	20	
Stone et. al (2009)[a]	4	59 dichotomous and polytomous items altogether for 4 subtests		10,545			√
Fu & Qu (2010)	2	a combination of number of in-scale items (5, 10, 20, and 30) and number of out-scale items(5, 10, 20, and 30)	0.1, 0.5, & 0.9	2,000	√	50 or 100	
Skorupski & Carvajal (2010)	4	52 items for 4 subtests (15, 12, 14 and 11 respectively)		17,226			√
de la Torre (2011)[a]	2 &5	10, 20, & 30 for each subtest	0, 0.4, 0.7 & 0.9	1,000	√		
	4	90 altogether for 4 subtests (25, 20, 20, and 25 respectively)	About 0.75 averagely	2,255			√

Note. [a] For the sake of space in the table, only the first author in their study is listed. Please check the reference list for detailed information. [b]The ratio of constructed-responses (CR) items to multiple-choice (MC) items was also considered as a factor in the study.

lengths, correlations between subscales, and sample sizes. By comparison, a vast number of studies, not limited to the ones cited in Tables 1 and 2, provided sound and solid evidence on the advantages of these improved subscoring methods. By examining the methods in different testing conditions, some valuable guidelines were also addressed as future references in the studies.

As a matter of fact, regarding the critical challenges of subscore reporting, Edwards and Vevea declared three possible solutions in their study in 2006, which were (1) to increase the subtest length, (2) to adopt collateral information derived among subtests, and (3) to consider the computerized adaptive testing (CAT) for customized test assembling. The subscoring methods mentioned above led to their second solution and have achieved the intended purposes to a large extent. In terms of the other two solutions, the first one is unrealistic because time and testing resources are limited and total scores are always of the most importance and interest for the majority of assessments. Adding more items in each subtest may supply redundant information when total scores are estimated and also increase the undesirable testing time and cost.

In the recent decades, the last solution, also suggested by Wainer et al. (2001), becomes more and more promising and feasible because many large-scale assessments gradually adopt computerized adaptive testing (CAT) as their testing format with the aid of advanced testing and computer technology. As widely recognized, the most advantageous characteristic of CAT, compared to the conventional P&P tests, is that it assembles a real-time test tailored by the just-in-time performance of examinees in the course of a test, and provides relatively more accurate and reliable ability parameter estimates right after the test is completed. Moreover, under the condition of ensuring the comparable accuracy and reliability of estimates, CAT requires a shorter test length than the P&P tests, which is even applicable to the examinees with extreme abilities if item pools are fully constructed. Considering this advantage, CAT may potentially provide a resolution to the less accurate and reliable subscore estimates that always result from insufficient items in the subscoring procedure.

In addition, formal assessments for Common Core Standards (2010) are expected to launch during the 2014-2015 school year. One of the testing formats is adaptive online tests. By then, in order to meet the standards and enhance the readiness of high school graduates for the future, it is foreseeable that CAT subscoring mechanism will be in great demand from participating states for its diagnostic values. In the meantime, test developers must be aware that additional assessments particularly designed for diagnostic purposes are not very adoptable in practice considering the incremental testing frequency and expenses. The optimal alternative therefore turns to the possibility of pulling the diagnostic information out of the conventional large-scale assessments as well as maintaining the original test purposes and specifications. In other words, attempts should be made to figure out some approaches of deriving both total scores and subscores from the same large-scale assessments at one time and simultaneously achieving the desirable accuracy and reliability of both types of scores. For the subscoring methods listed in Table 1, only Yen’s (1987) and de la Torre & Song’s (2009) methods can provide both scores at the same time in one test.

Besides, another concern points to the fact that all the subscoring approaches mentioned previously were developed on the traditional P&P testing format. Their extensions and applications in CAT could be very desirable. With the growing popularity of CAT, the MIRT model has been expanded to the CAT framework and is correspondingly developed as the multidimensional adaptive testing (MCAT, Segall, 1996). Moreover, given the specific operational features of CAT, van der Linden (2010) proposed an estimation algorithm to improve subscore estimates by adaptively sequencing subtests in a test battery (SEQ-CAT). Recently, he continued developing this algorithm by maximally utilizing the information derived from the complete response pattern and correlation structure (reSEQ-CAT; W. J. van der Linden, personal communication, July 30th, 2013). Through the investigations of some studies, these computer-based adaptive scoring methods have been identified as more efficient and reliable score estimation approaches, compared to the conventional unidimensional adaptive test (UCAT) scoring and multiple independent UCAT (IND-UCAT) scoring (Luecht, 1996; Segall, 2001; Li & Schafer, 2005; Yao, 2012; van der Linden, 1999; van der Linden, 2010).

In addition, Luo, Diao, and Ren (2014) applied Wainer's AUG to the simulated CAT tests (AUG-CAT). As one of the most widely accepted subscoring methods in P&P tests, the augmentation technique developed in Wainer's AUG indubitably deserves special attentions and endeavors as it is combined with the adaptive testing algorithm. More importantly, it demands for relatively simpler computation compared to MCAT, SEQ-CAT and reSEQ-CAT, and consequently might be more applicable to the operational tests if it could ensure the quality of subscore estimates as the other three methods do. Therefore, aside from an interest in the application of Wainer's AUG in CAT tests, comparing it with the other three CAT subscoring methods is also worth considerable attentions, which, to date, has not yet been presented in the literature.

As another widely-recognized subscoring method in P&P tests, Yen's OPI seems very promising to be developed to the CAT and also indispensable to compete with the other CAT subscoring methods in the study. However, under the CAT framework, OPI's original design constrains its application to CAT tests. More precisely, Yen's OPI in P&P tests is defined as the mean of the posterior distribution of the true proportion-correct subscore $\pi_{i(d)}$, which is estimated by $\frac{1}{J_{(d)}}\sum_{j_{(d)}=1}^{J_{(d)}} P_{ij_{(d)}}(\hat{\theta}_i)$. $P_{ij_{(d)}}(\hat{\theta}_i)$ is the probability of a correct response to item *j* in subtest *d* for examinee *i* with the general ability estimate of $\hat{\theta}_i$, and $J_{(d)}$ is the test length of subtest *d*. In a P&P test, all the items in each subtest completed by individual examinees are fixed and identical. Therefore, the prior proportion-correct subscore estimates $\tilde{\pi}_{i(d)}$ are comparable among all examinees. However, in a CAT test, the items optimally measuring the real-time ability estimate are adaptively selected from the item pool for individual examinees. That is, the items selected for each examinee might be very different depending on their just-in-time performance during the test. Given this characteristic of CAT, the use of the proportion-correct subscores to distinguish examinees is totally inappropriate because the probability of a correct response to an item that matches the provisional ability estimate in CAT is always approximate to 0.50 regardless of the placement of examinees on the ability scale. Considering this limitation, Yen's OPI is not included in the study.

Purpose of the Study

Based on the discussion above, the primary objective of the study is to compare some CAT subscoring methods by evaluating their subscore estimation on dichotomously-scored items in CAT tests, as has been conducted in the P&P tests in the literature. The subscoring methods mentioned above are considered for comparison. However, in order to make the comparisons more comparable and realistic, the study modifies the conventional MCAT as the pool-constrained MCAT (PC-MCAT), which is described in more detail in the 2nd section of Chapter 3. Namely, the study includes IND-UCAT, AUG-CAT, SEQ-CAT, reSEQ-CAT, and PC-MCAT in the comparisons, in which IND-UCAT is treated as the baseline subscoring method. Some relevant factors listed in Table 2 for P&P tests are also crucial to CAT tests, and therefore their effects on CAT subscore estimation are worth investigating. Two of the factors, subtest length and the correlations between subtests, are considered in the study.

Also, in most large-scale assessments, each subtest usually measures a particular subscale or content area, which implies a simple structure that each item loads only on one subscale or content domain. In the study, all the items are derived from real existing subpools and they all exhibit a simple structure. In addition, as the methods originally designed for CAT tests, IND-UCAT, SEQ-CAT, and PC-MCAT have their own item selection algorithms, which demonstrate different capacities of exploiting the collateral information in the item selection procedure and may agitate the comparability of these scoring methods. Therefore, the three item selection algorithms are also taken into consideration in the study, and are individually conducted along with all the five scoring methods. As a byproduct of this consideration, the performances of these three item selection algorithms on improving score estimates are also demonstrated in the study.

Furthermore, as indicated in Wainer et.al's study (2001), tests are most commonly used for ranking and diagnosis. In practice, most large-scale assessments are designed merely for the first purpose, aiming to seek the standings of examinees on a common general ability scale. To ensure the validity and fairness of ranking, the assessments need to cover a wide range of contents or subscales within a subject or a general ability to align with test specifications and also to avoid favoring certain groups of examinees. Recall that a wide coverage of a test on contents or subscales can lead to inadequate items in each subtest, and thus prompts big challenges for subscore estimation that is typically the derivation of diagnostic information. However, given the increasing voices for subscore reports in the market, large-scale standardized tests are imperatively expected to hold capabilities of serving for both purposes at no expense of testing cost and time. Under the circumstance, this study takes advantage of the hierarchical latent trait structure and suggests an approach of calculating the total scores based on the subscores estimated by the subscoring methods described. This approach is applicable to both P&P and CAT testing formats, of which the latter is the focus of the study, and provides subscore and total score estimates successively from one test.

Research Questions

To accomplish the purposes of the study, a simulation study is designed to mimic different CAT testing conditions. The following five aspects of research questions are addressed in the study:

1. How well do the other four subscoring methods perform in improving the accuracy of subscore estimates under various CAT testing conditions compared to the baseline method of the multiple independent UCAT (IND-UCAT) scoring procedure?
2. How comparatively efficient are the other four subscoring methods in subscore estimation under various CAT testing conditions other than IND-UCAT?
3. How do the investigated factors, including subtest length and the correlations between subtests, influence the performance of the five subscoring methods?
4. How well does the suggested successive scoring approach perform in recovering the true total scores under various CAT testing conditions?
5. How well do the three item selection algorithms perform under various CAT testing conditions? Which combination of the scoring method and the item selection method performs the best under the conditions?

CHAPTER II

LITERATURE REVIEW

The discussion above, especially the summary in Table 1, exhibits some existing subscore estimation approaches for P&P and CAT testing formats in the modern assessment realm. In the last few decades, a number of the P&P subscoring methods have been widely accepted, thoroughly compared and some even applied to the real P&P tests by measurement researchers and practitioners. For the study, five primary subscoring methods fitting in the CAT testing environments are compared. Correspondingly, the studies regarding their rationale, applications and comparisons are theoretically and technically described in this chapter. Before jumping to the details of these five subscoring methods, some components relevant to the implementation of a CAT test are first introduced, which are employed across all the five subscoring methods. They primarily include the types of estimated IRT ability parameters, item selection criteria, and the methods of constraint imposition. For comparison purposes, the use of the consistent components across all methods is prerequisite and vital. In addition, the higher-order IRT model is briefly described for a reason that the structural phase in HO-IRT model provides a clue for the study to calculate the general ability scores.

Maximum A Posterior (MAP) Estimates

Maximum a posterior (MAP, see Samejima, 1969) estimates are developed from Bayesian estimation philosophy, for which the ability parameter estimation is relatively precise, efficient and feasible, especially when higher ability dimensionality and extreme response patterns are involved (e.g. Bock & Aitken, 1981; Bock & Mislevy, 1982; Lord, 1986, Swaminathan & Gifford, 1986; Segall, 1996; Chen, 2009). Conceptually speaking, the Bayesian estimation is implemented by incorporating the previous knowledge (the prior distribution) into the data analysis process (the likelihood function) to shape the new evidence (the posterior density function) on the target parameters. It constantly updates the beliefs on all the uncertain quantities including unobserved parameters by utilizing the newly-input information from the data. One of the advantages of Bayesian estimation is that unobserved parameters that might be poorly estimated based on the data can be improved in conjunction with the proper informative prior distribution. The prior distribution is often elicited from modeling the previous studies and the beliefs from experts. The posterior density function is, by definition, expressed as

$$f(\theta_{(d)} \mid \boldsymbol{u}_{i(d)}) = \frac{L(\boldsymbol{u}_{i(d)} \mid \theta_{(d)})\, f(\theta_{(d)})}{f(\boldsymbol{u}_{i(d)})} \qquad (1)$$

$$f(\boldsymbol{\theta} \mid \boldsymbol{u}_i) = \frac{L(\boldsymbol{u}_i \mid \boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(\boldsymbol{u}_i)}, \qquad (2)$$

in which, for the unidimensional IRT case (Equation 1), $\theta_{(d)}$ refers to the ability parameter in subtest *d* while for the multidimensional IRT case (Equation 2), $\boldsymbol{\theta}$ refers to a vector of ability parameters in all subtests; $\boldsymbol{u}_{i(d)}$ or $\boldsymbol{u}_i$ respectively represents the observed response pattern of examinee *i* in subtest *d* or in an entire test; $L(\boldsymbol{u}_{i(d)} \mid \theta_{(d)})$ or $L(\boldsymbol{u}_i \mid \boldsymbol{\theta})$ is respectively the likelihood function of the observed responses $\boldsymbol{u}_{i(d)}$ or $\boldsymbol{u}_i$;

$f(\theta_{(d)})$ is the prior distribution of $\theta_{(d)}$ while $f(\boldsymbol{\theta})$ is the multivariate prior distribution of $\boldsymbol{\theta}$; and $f(\boldsymbol{u}_{i(d)})$ or $f(\boldsymbol{u}_i)$ is respectively the marginal probability density function of $\boldsymbol{u}_{i(d)}$ or $\boldsymbol{u}_i$, which plays as a constant for normalization.

Based on the assumption of local independence, the likelihood functions in Equations (1) and (2) for the first *k* -1 items administered in CAT are accordingly defined as

$$L(\boldsymbol{u}_{i(d)}^{k-1} \mid \theta_{(d)}) = \prod_{j_{(d)}=1}^{k-1} P_{j_{(d)}}(\theta_{(d)})^{u_{ij_{(d)}}} (1 - P_{j_{(d)}}(\theta_{(d)}))^{1-u_{ij_{(d)}}} \tag{3}$$

$$L(\boldsymbol{u}_i^{k-1} \mid \boldsymbol{\theta}) = \prod_{j=1}^{k-1} P_j(\boldsymbol{\theta})^{u_{ij}} (1 - P_j(\boldsymbol{\theta}))^{1-u_{ij}} \quad , \tag{4}$$

where $P_{j_{(d)}}(\theta_{(d)})$ or $P_j(\boldsymbol{\theta})$ is respectively the probability of a correct response to item *j* in subtest *d* or in a test measuring multiple abilities; $\boldsymbol{u}_{i(d)}^{k-1}$ or $\boldsymbol{u}_i^{k-1}$ represents the response pattern to the first *k*-1 items administered in subtest *d* or in a test for examinee *i*; and $u_{ij_{(d)}}$ or u_{ij} is the response to item *j* in subtest *d* or in a test for examinee *i*.

In general, MAP estimates, also known as Bayes modal estimates (BMEs), are the values of ability parameters corresponding to the maximum point of the posterior density function. They occur when the first derivative(s) of the posterior density function is (are) equal to 0. For the sake of computational convenience, the natural logarithm of the posterior distribution is usually used. In the unidimensional CAT case, the MAP subscore $\hat{\theta}_{i(d)}^{k-1}$ for examinee *i* estimated from the first *k*-1 selected items in subtest *d* is known as

$$\hat{\theta}_{i(d)}^{k-1} = \arg\max_{\theta_{i(d)}}\{\ln f(\theta_{(d)} \mid \boldsymbol{u}_{i(d)}^{k-1})\}, \tag{5}$$

in which

$$f(\theta_{(d)} \mid \boldsymbol{u}_{i(d)}^{k-1}) = \frac{L(\boldsymbol{u}_{i(d)}^{k-1} \mid \theta_{(d)}) f(\theta_{(d)})}{f(\boldsymbol{u}_{i(d)}^{k-1})} \tag{6}$$

and is the updated posterior distribution of $\theta_{(d)}$ after counting in the (*k*-1)th response in subtest *d*. By Equation (5), $\hat{\theta}_{i(d)}^{k-1}$ is approximately obtained by

$$\frac{\partial}{\partial\theta_{(d)}} \ln f(\theta_{(d)} \mid \boldsymbol{u}_{i(d)}^{k-1}) = 0. \tag{7}$$

In the multidimensional CAT case, the MAP subscores $\hat{\boldsymbol{\theta}}_i^{k-1}$ for examinee *i*, estimated from the first *k*-1 selected items in a test, are the approximation to the IRT scale scores that maximize the natural logarithm of the posterior density function $f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1})$ in the multiple-dimensional space (Segall, 1996). That is,

$$\hat{\boldsymbol{\theta}}_i^{k-1} = \arg\max_{\boldsymbol{\theta}_i}\{\ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1})\}. \tag{8}$$

Mathematically, they are the solutions to a set of *D* simultaneous equations

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) = 0, \tag{9}$$

where

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) = \begin{bmatrix} \dfrac{\partial}{\partial \theta_{(1)}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) \\ \dfrac{\partial}{\partial \theta_{(2)}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) \\ \vdots \\ \dfrac{\partial}{\partial \theta_{(D)}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) \end{bmatrix} \tag{10}$$

and *D* is the total number of the subscales measured by a test. In the study, because all the items exhibit a simple structure, *D* is also the total number of subtests included in the test battery. The individual partial derivative with respect to each subscale in Equation (10) could be further expressed as

$$\frac{\partial}{\partial \theta_{(d)}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1}) = \frac{\partial}{\partial \theta_{(d)}} \ln L(\boldsymbol{u}_i^{k-1} \mid \boldsymbol{\theta}) + \frac{\partial}{\partial \theta_{(d)}} \ln f(\boldsymbol{\theta}). \tag{11}$$

Because there are no closed form solutions to Equations (11), some iterative numerical procedure is required. Suppose that $\hat{\boldsymbol{\theta}}^{(m)}$ represents the *m*th approximation to the values of $\boldsymbol{\theta}$ that maximize $\ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1})$. Then the next approximation $\hat{\boldsymbol{\theta}}^{(m+1)}$, which could produce larger $\ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i^{k-1})$, is given by

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\delta}^{(m)}, \tag{12}$$

in which $\boldsymbol{\delta}^{(m)}$ is a $D\times 1$ vector, denoted as

$$\boldsymbol{\delta}^{(m)} = [\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) \,. \tag{13}$$

For the Newton-Raphson iterative procedure, $\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})$ in Equation (13), known as Hessian matrix, is a $D\times D$ symmetric matrix with elements of second derivatives evaluated at $\hat{\boldsymbol{\theta}}^{(m)}$, which is

$$\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)}) = \frac{\partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1})}{\partial \theta \partial \theta'} \,. \tag{14}$$

The elements in $\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})$ are more specifically written as

$$\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)}) = \begin{bmatrix} \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_1^2 & \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_1 \partial \theta_2 & \cdots & \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_1 \partial \theta_D \\ & \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_2^2 & \cdots & \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_2 \partial \theta_D \\ & & \ddots & \vdots \\ & & & \partial^2 \ln f(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{u}_i^{k-1}) / \partial \theta_D^2 \end{bmatrix} .$$

By Equations (12) and (13), the approximation process is repeated until the elements in $\boldsymbol{\delta}^{(m)}$ become very small. The final approximation is accordingly treated as the MAP subscore estimates $\hat{\boldsymbol{\theta}}_i^{k-1}$ for examinee *i* after taking the first *k*-1 items. Sometimes, the iterative procedure may not be converged when the selection of the initial values of $\hat{\boldsymbol{\theta}}^{(m)}$ does not fall near the true global maximum. For such a situation, Segall (1996, 2010) suggested to use Fisher's scoring method to avoid non-convergence, which is to replace $\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})$ in Equation (13) by

$$\mathrm{E}[\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})] = -\boldsymbol{I}^{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(m)}), \tag{15}$$

where $\boldsymbol{I}^{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(m)})$ is negative the expected $\boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(m)})$ and known as the Fisher's posterior information matrix evaluated at $\hat{\boldsymbol{\theta}}^{(m)}$, which is described in detail in the section of MCAT of this chapter.

After obtaining the MAP estimates $\hat{\boldsymbol{\theta}}_i^{k-1}$ based on the first *k*-1 items, the test proceeds to consecutively select the next few optimal items until some stopping rule is satisfied. The values of $\hat{\boldsymbol{\theta}}_i$ that maximize the last updated posterior density function $\ln f(\boldsymbol{\theta} \mid \boldsymbol{u}_i)$ are regarded as the ultimate MAP subscore estimates. As a matter of fact, the iterative procedure used in the multidimensional CAT can also be applied to the unidimensional CAT such that the solution to Equation (7) can be faster obtained. The Newton-Raphson procedure in UCAT is demonstrated as

$$\theta_{(d)}^{(m+1)} = \theta_{(d)}^{(m)} - \delta_{(d)}^{(m)}, \tag{16}$$

where

$$\delta_{(d)}^{(m)} = \frac{\partial \ln f(\theta_{(d)}^{(m)} \mid \boldsymbol{u}_{i(d)}^{k-1}) \big/ \partial \theta_{(d)}}{\partial^{2} \ln f(\theta_{(d)}^{(m)} \mid \boldsymbol{u}_{i(d)}^{k-1}) \big/ \partial \theta_{(d)}^{2}}. \tag{17}$$

In this study, the MAP ability estimation is adopted primarily for two reasons. First, compared to maximum likelihood estimates (MLE), IRT Bayesian ability score estimates, mainly referred to as MAP estimates and expected a posterior (EAP) estimates, are obtainable for the examinees with extreme response patterns, which include the null and perfect response patterns. Also, the Bayesian estimates are relatively more precise and efficient for the fact that they yield lower standard error (SE) in CAT tests, especially in short tests (Warm, 1989; Wang & Vispoel, 1998). As discussed previously, insufficient items used for subscoring in each subtest may lead to unreliability and imprecision of subscore estimates. Under the framework of Bayesian estimation, the insufficiency of items can be somewhat compensated by adding the prior knowledge on the (multivariate) distribution of the ability (abilities) on the subscale (subscales). Second, the MAP estimation demands much less computation than the EAP estimation when a large number of subscales are involved in a CAT test and thus turns out to be more feasible for some computer programs (Segall, 1996; Chen, 2009). For EAP estimation, the quadrature points are often used to obtain the approximation to the integration. If thirty quadrature points are applied to each ability dimension, multiple combinations of thirty points across dimensions could exponentially increase as the number of dimensions increases. Therefore, as the number of subscales is large, the time for computation would be a very critical issue, which largely impairs the advantages of CAT in practice.

Maximum Posterior-Weighted Information (MPI) For Item Selection in UCAT

Over the conventional linear tests, a major advantage of CAT is the real-time item selection, which indicates a procedure of searching the following item that optimally measures the current ability score estimate. That is, as the ability score estimate is updated, only the items right tailored for individual examinees enter the test. A variety of item selection criteria are developed for the framework of UCAT, which primarily include the maximum-information criterion (Weiss, 1982), the Bayesian criterion (van der Linden, 1998) and the Kullback-Leibler information criterion (Chang & Ying, 1996). Distinguished from Owen's (1975) approximate Bayesian criterion, van der Linden's (1998) Bayesian criteria is a fully Bayesian procedure that implements item selection based on the full posterior, and is mainly referred to as the criteria of maximum posterior-weighted information (MPI), maximum expected information, minimum expected posterior variance, and maximum expected posterior weighted information (van der Linden, 1998). The first approach, MPI, is used as the item selection criterion for all the UCAT subscoring methods in the current study, which also aligns with the original design of SEQ-CAT (van der Linden, 2010).

Maximum posterior-weighted information (MPI) criterion is essentially a reformulation of the maximum information criterion within the framework of Bayesian inferences, which is an algorithm of seeking the following item with the maximum expected information over the posterior distribution. This item selection criterion allows for the integration of empirical data and the updated knowledge of the posterior distribution. Moreover, it permits the inclusion of the neighboring ability score estimates yielding considerable likelihoods in the course of item selection.

For instance, in terms of MAP scale scores, as the (*k*-1)th item in subtest *d* is completed, the new response is thereafter used to update the posterior distribution $f(\theta_{(d)} \,|\, \boldsymbol{u}_{i(d)}^{k-1})$ by Equation (6) and then the new MAP estimate $\hat{\theta}_{i(d)}^{k-1}$ is obtained. For the MAP ability scores employed in the study, the corresponding Fisher's posterior information function $I^{p}(\theta_{(d)}, \hat{\theta}_{(d)}^{k-1})$ is denoted as

$$\begin{aligned} I^{p}(\theta_{(d)},\hat{\theta}_{(d)}^{k-1}) &= -\frac{\partial^{2}}{\partial\theta_{(d)}^{2}}\ln L(\boldsymbol{u}_{i(d)}^{k-1}\mid\theta_{(d)})\,f(\theta_{(d)}) \\ &= \sum_{j_{(d)}=1}^{J_{(d)}}\frac{[P_{j_{(d)}}^{'}(\theta_{(d)})]^{2}}{[P_{j_{(d)}}(\theta_{(d)})][1-P_{j_{(d)}}(\theta_{i(d)})]}+\sigma^{-2}(\theta_{(d)}), \\ &= I(\theta_{(d)},\hat{\theta}_{(d)}^{k-1})+\sigma^{-2}(\theta_{(d)}) \end{aligned} \tag{18}$$

where $I(\theta_{(d)},\hat{\theta}_{(d)}^{k-1})$ is the Fisher's information function regarding maximum-likelihood estimates (MLEs); $P_{j_{(d)}}^{'}(\theta_{(d)})$ is the first derivative of $P_{j_{(d)}}(\theta_{(d)})$; and $\sigma^{-2}(\theta_{(d)})$ is the reciprocal of the variance of the prior distribution. Because the second term on the right of Equation (18) is constant for all items in subpool *d*, the Fisher's information function, regardless of the types of UIRT scale scores, is considered when selecting the next item in a CAT test, as opposed to the Fisher's posterior information function.

As for the maximum-information criterion, the Fisher's information function at a single point estimate of the ability parameter (say, $\hat{\theta}_{i(d)}^{k-1}$) is the only determinant for the next-item selection. Specifically speaking, when selecting the *k*th item in subtest *d* from the remaining of subpool $R_{k(d)}$, the item that maximizes the Fisher's information function evaluated at $\hat{\theta}_{i(d)}^{k-1}$ would be selected. That is,

$$j_{ik(d)} = \arg\max_{j_{ik(d)}}\{I_{k'}(\theta_{i(d)},\hat{\theta}_{i(d)}^{k-1});\, j_{ik(d)}\in R_{k(d)}\}\,. \tag{19}$$

Based on the assumption of conditional independence given $\theta_{i(d)}$, the Fisher's information function is additive. By notation, that is

$$I(\theta_{i(d)}, \hat{\theta}^{k}_{i(d)}) = I(\theta_{i(d)}, \hat{\theta}^{k-1}_{i(d)}) + I(\theta_{i(d)}, u_{ik(d)}) \, . \tag{20}$$

Because the first term on the right of Equation (20) holds constant for $R_{k(d)}$, Equation (19) is equivalent to

$$j_{ik(d)} = \arg\max_{j_{ik(d)}} \{ I_{k'}(\hat{\theta}^{k-1}_{i(d)}, u_{ik'(d)}); \, j_{ik(d)} \in R_{k(d)} \} \, , \tag{21}$$

where $I_{k'}(\hat{\theta}^{k-1}_{i(d)}, u_{ik'(d)})$ is the information of the candidate of the *k*th item evaluated at $\hat{\theta}^{k-1}_{i(d)}$.

Regarding the MPI criterion, when the *k*th item in subtest *d* is to be selected from $R_{k(d)}$, the item that maximizes the expected Fisher's information over the updated posterior distribution would be selected. It is denoted as

$$j_{ik(d)} = \arg\max_{j_{ik(d)}} \{ \int I_{k'}(\theta_{(d)}, u_{k'(d)}) \, f(\theta_{(d)} \mid \boldsymbol{u}^{k-1}_{i(d)}) d\theta_{(d)}; \, j_{ik(d)} \in R_{k(d)} \} \, , \tag{22}$$

which demonstrates that the information produced by any candidate item from $R_{k(d)}$ is weighed by the posterior distribution of $\theta_{(d)}$. The weights are a function of the likelihood and the prior distribution over the entire ability scale continuum. It implies that the item that optimally measures a narrow ability interval rather than an ability point estimate is most likely selected as the next item by the MPI criterion. The considerations on the other likely ability points in the neighborhood can be of great benefit to efficiently select the items that most likely match the true ability score. By contrast, the maximum-information criterion for MAP scores simply considers the item with the largest maximum information evaluated at a single ability point estimate. However, at the early stage of a test, the likelihood function is typically flat and has less impact on the posterior distribution. As a result, the posterior distribution will be very approximate to the initial prior distribution, so no significant differentiation is expected between the new MAP score estimate and the initial ability score. Under such a circumstance, selecting an item with maximum information evaluated at a single ability estimate may slow down the posterior distribution converged at the true point.

Shadow Test

As discussed above, Fisher's information function plays an important role in CAT item selection. However, the item selection procedure simply depending on the information function may, in practice, result in some nonstatistical violations of test specifications, such as unbalanced content areas, disproportional answer keys, or item over-exposure. In order to ensure test specifications, some pertinent constraints are always imposed during the process of selecting each item. With the constraints imposed, the selected item needs simultaneously to guarantee the maximization of statistical information. The algorithm accomplishing both goals was named constrained sequential optimization in van der Linden's study (2010). Prior to his study, some methods had been developed to implement the constrained sequential optimization, which involved maximum priority index method (Cheng & Chang, 2009), item-pool partitioning (Kingsbury & Zara, 1991), weighted-deviation method (Swanson & Stacking, 1993), and multistage testing (Adema, 1990), etc. However, the results from the investigations on these methods showed that these methods might lead to a dilemma, either violations of some constraints or suboptimal adaptation at the end of a test (van der Linden, 2005).

The shadow test proposed by van der Linden and Reese (1998) breaks through the dilemma and fulfills the optimal adaptation and the realization of all constraints simultaneously. Also, note that the shadow test is not a common-sense test for administration, but an algorithm of a real-time test assembly. It starts with assembling a full-length test (the first shadow test) that includes the first few items with the maximum information at the initial ability estimate, under the condition that all the constraints are satisfied. Then the item providing the maximum information is selected from the first shadow test, instead of from the item pool, and is administered. Thereafter the ability estimate is updated, and then a new shadow test is correspondingly assembled not only with both goals achieved but also with the earlier administered item included. It continues until some stopping rule is satisfied. In the current study, the last shadow test would contain all the actually administered items and simultaneously meet all the constraints.

In principle, the shadow test is a composition of the maximization of statistical information and the realization of nonstatistical specifications (van der Linden, 2010). It is implemented by maximizing the objective function with decision variables manipulated, so that the constraints depending on the test specifications could be imposed and the eligible items with maximum information could be selected into the shadow test. For instance, to make the subscores estimated by different methods comparable, the current study employs the fixed subtest length across the compared methods (and therefore the total test length is also fixed). To meet this specification, in the case of UCAT with MPI criterion and MAP scores, when the *k*th item is selected for subtest *d*, the objective function for the *k*th shadow test is expressed as

$$\text{maximize}\sum_{n_{(d)}=1}^{N_{(d)}}\left[\int I(\theta_{(d)},u_{n_{(d)}})f(\theta_{(d)}\mid \boldsymbol{u}_{i(d)}^{k-1})d\theta_{(d)}\right]x_{n_{(d)}} \qquad (23)$$

subject to

$$\sum_{n_{(d)}=1}^{N_{(d)}} x_{n_{(d)}} = J_{(d)} \qquad (24)$$

$$x_{n_{(d)}} = 1, \textit{ for all } n_{(d)} \in S_{(k-1)_{(d)}} \qquad (25)$$

$$x_{n_{(d)}} \in \{0,1\},\ n_{(d)} = 1,2,...,N_{(d)}, \qquad (26)$$

where $N_{(d)}$ is the total number of items in subpool d; $J_{(d)}$ is the test length of subtest *d*; $x_{n_{(d)}}$ is the binary decision variable for the selection of item $n_{(d)}, n_{(d)} = 1,2,...,N_{(d)}$; and $S_{(k-1)_{(d)}}$ is the set of the first (*k*-1) selected items in the shadow test from subpool *d*.

As such, with the sublength constrained in MCAT that adopts the item selection criterion of the Bayesian version of D-optimality (for more details, see the section of MCAT of this chapter), when the *k*th item is selected, the objective function for the *k*th shadow test is expressed as

$$\text{maximize}\sum_{n=1}^{N}\left[\det(\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k-1})+\boldsymbol{I}(\hat{\boldsymbol{\theta}}^{k-1},u_{in})+\boldsymbol{\Phi}^{-1})\right]x_n \qquad (27)$$

subject to

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{J} \qquad (28)$$

$$x_n = 1, \textit{ for all } \; n \in S_{(k-1)} \tag{29}$$

$$x_n \in \{0,1\}, \; n = 1, 2, ..., N \,, \tag{30}$$

where *N* is the total number of items in the item pool and $N = \sum_{d=1}^{D} N_{(d)}$; $\mathbf{A}$ is a $D \times N$ matrix with binary elements of either 0 or 1, reflecting the mapping of all items in the entire pool; $\boldsymbol{x}$ is a vector including the binary decision variables for the selection of items from 1 to *N*; $\boldsymbol{J}$ is a vector with elements of $J_{(1)}, J_{(2)}, ..., J_{(D)}$, $J_{(1)} = J_{(2)} = ... = J_{(D)}$ in the study, and $J = \sum_{d=1}^{D} J_{(d)}$, where J is the total number of items administered in a MCAT test.

Equations (24) to (26) and Equations (28) to (30) represent the constraints requested in the test specifications in the current study. The numbers of items from each subscale are constrained by Equations (24) and (28). Another constraint of Equations (25) and (29) indicates the inclusion of all the *k*-1 items that have been administered into the *k*th shadow test. The decision variable $x_{n_{(d)}}$ or x_n is constrained as a binary variable of either 0 or 1 by Equations (26) and (30), in which item $n_{(d)}$ or *n* is selected into the shadow test if $x_{n_{(d)}}$ or x_n is equal to 1 and otherwise item $n_{(d)}$ or *n* is not selected. Basically, how to determine the values of the decision variables becomes the core of accomplishing the constrained sequential optimization problem in a shadow test. Their proper values should simultaneously solve equations (24) through (26) in UCAT or equations (28) through (30) in MCAT. Just as van der Linden stated, “A solution to the optimization problem is a vector of zeros and ones for the decision variables that identifies the set of items that meets the constraints and has a maximum value for the objective function" (2010, p.108).

To locate the values of decision variables, a 0-1 integer linear programming (ILP) is always adopted. A powerful solver to the ILP model can efficiently find out the solutions based on branch-and-bound (BAB) or some other methods. In addition, aside from the constraints mentioned above, some other categorical, quantitative, or even logical attributes of an item can also be constrained, such as word counts, item format, or enemy items. Imposing other constraints in ILP models is discussed in more detail by van der Linden (2010).

Wainer's Augmented Subscoring (AUG) and AUG-CAT

Wainer's augmented subscoring procedure (AUG, see Wainer et al, 2001) is a regression-based empirical Bayes subscore estimation approach, currently applied only to the P&P operational testing format. Its augmentation algorithm behaves like a multiple regression of a true subscore on all the observed deviation subscores in a test. The observed group mean of a target subtest is included as the intercept in the regression function. The regression coefficients are largely determined by the reliabilities of subtests and their correlations to the target subtest. That is, the subtests with high reliabilities and high correlations to the target subtest are more likely to be granted larger weights on estimating the true target subscore. Thus, a subscore estimate is augmented by exploiting the collateral information from all the other subtests, other than simply depending on the information observed within one subtest.

Wainer's AUG was originally derived from Kelley's (1927, 1947) regressed equation for a true score, which is shown as

$$\hat{\tau} = \rho_{xx'} x + (1 - \rho_{xx'})\mu , \tag{31}$$

where the augmented true score estimate $\hat{\tau}$ is calculated by regressing the observed score *x* toward the test group mean μ to an extent depending on the magnitude of the test reliability $\rho_{xx'}$. The test reliability $\rho_{xx'}$ is estimated by S_{true}^2 / S_{obs}^2, in which S_{true}^2 and S_{obs}^2 are respectively the estimated true variance and the observed variance from the sample. Kelley's regressed equation can also be rewritten as

$$\hat{\tau} = \mu + \rho_{xx'}(x - \mu) . \tag{32}$$

Considering a test composed of a test battery, Wainer et al. (2001) generalized Kelley's equations to the multivariate form under the same assumption that true scores and observed scores all follow a (multivariate) normal distribution. By sample notation, Wainer's multivariate regressed equation is expressed as

$$\begin{aligned} \hat{\boldsymbol{\tau}} &= \boldsymbol{Bx} + (1 - \boldsymbol{B})\overline{\boldsymbol{X}} \\ &= \overline{\boldsymbol{X}} + \boldsymbol{B}(\boldsymbol{x} - \overline{\boldsymbol{X}}) \end{aligned} , \tag{33}$$

where $\hat{\boldsymbol{\tau}}$ is a vector of augmented true subscore estimates; $\overline{\boldsymbol{X}}$ and $\boldsymbol{x}$ are the vectors of subtest means and observed subscores; and ***B*** is the reliability-related coefficient matrix.

The coefficient matrix ***B*** contains the weights for all the linear combinations of deviation subscores. These linear combinations are actually the equations of estimating all the true target subscores. The matrix $\boldsymbol{B}$ is determined by the reliabilities and intercorrelations of subtests and is calculated by $\boldsymbol{S}_{true}\boldsymbol{S}_{obs}^{-1}$. In alignment with the counterpart of reliability S_{true}^2 / S_{obs}^2 in the univariate case, $\boldsymbol{S}_{true}$ and $\boldsymbol{S}_{obs}$ are respectively the estimated true covariance matrix and the observed covariance matrix from the sample. When $\boldsymbol{B}$ is an identity matrix implying perfect reliability and independence of subscores, Wainer's AUG estimates are reduced as observed subscores; When the subscores depart from perfect reliability implying the occurrence of measurement errors, the information contributed by the other subtests is added to the true target subscore estimation; When $\boldsymbol{B}$=$\boldsymbol{0}$ implying absolutely independent and unreliable subscores, Wainer's AUG estimates are reduced as subscore means.

By Equation (33), the augmentation procedure in Wainer's AUG is obviously demonstrated, which is to weigh the information from all the other subtests on the target subscore estimation, according to their reliabilities and the correlations to the target subtest. Wainer's augmented subscoring procedure is applicable to the classical observed scores and IRT scale scores. Since MAP estimates are used in the study, the following discussion focuses on the derivation of Wainer's augmented subscore estimates from IRT MAP scale scores.

As demonstrated in Section 1 of this chapter, MAP scale scores $\hat{\theta}_{i(d)}$, calculated by Equations (5) and (6), are already augmented by shrinking the likelihoods towards the priorly-known population mean. In order to calculate $\boldsymbol{S}_{obs}$ from the original unaugmented observed score estimates, the shrinkage toward the prior distribution mean should be removed from MAP scale scores. That is, MAP scale scores $\hat{\theta}_{i(d)}$ should be converted to the unaugmented IRT score estimates $\hat{\theta}^{*}_{i(d)}$. Under the assumptions that the measurement errors across all the MAP scale scores are constant and MAP scale scores within each subtest have a zero mean, the conversion equation could be established as

$$\hat{\theta}^{*}_{i(d)} = \hat{\theta}_{i(d)} / \rho_{(d)}, \tag{34}$$

in which $\rho_{(d)}$ is the reliability of subtest *d* and is defined as

$$\rho_{(d)} = \frac{\sigma^{2}(\hat{\theta}_{(d)})}{\sigma^{2}(\hat{\theta}_{(d)}) + \overline{\sigma^{2}(\hat{\theta}_{(d)} \mid \boldsymbol{u}_{(d)})}}, \tag{35}$$

where $\sigma^{2}(\hat{\theta}_{(d)})$ is the variance of MAP subscore estimates in subtest *d* and $\overline{\sigma^{2}(\hat{\theta}_{(d)} \mid \boldsymbol{u}_{(d)})}$ is the average of the posterior variances of MAP subscores in subtest *d*. Also, the conversion equation (34) is transformed from the equation of $\hat{\theta}_{i(d)} = \rho_{(d)}\hat{\theta}^{*}_{i(d)}$, which is analogous to Kelley’s regression equation (Equation (31)).

By means of Equations (34) and (35), $\hat{\theta}^{*}_{i(d)}$ is calculated for each examinee in each subtest and the matrix of $\boldsymbol{S}_{obs}$ with respect to all the values of $\hat{\theta}^{*}_{i(d)}$ could also be calculated. Then, the matrix of $\boldsymbol{S}_{true}$ for true subscores is estimated from

$$\boldsymbol{S}_{true} = \boldsymbol{S}_{obs} - \boldsymbol{D}, \tag{36}$$

where $\boldsymbol{D}$ is a diagonal matrix with the *d*th diagonal element as $(1-\rho_{(d)})s_{obs}^{dd}$ in which s_{obs}^{dd} is the observed sample variance of subtest *d*. Given $\boldsymbol{S}_{true}$ and $\boldsymbol{S}_{obs}$, the matrix $\boldsymbol{B}^{*}$ for unaugmented IRT score estimates is obtained from $\boldsymbol{S}_{true}\boldsymbol{S}_{obs}^{-1}$. According to Equation (33), Wainer's MAP augmented subscore estimates are, therefore, given by

$$\begin{aligned}\hat{\boldsymbol{\theta}}_i^{AUG} &= \boldsymbol{B}^{*}\hat{\boldsymbol{\theta}}_i^{*} + (1-\boldsymbol{B}^{*})\overline{\hat{\boldsymbol{\theta}}^{*}} \\ &= \overline{\hat{\boldsymbol{\theta}}^{*}} + \boldsymbol{B}^{*}(\hat{\boldsymbol{\theta}}_i^{*} - \overline{\hat{\boldsymbol{\theta}}^{*}})\end{aligned}, \qquad (37)$$

where $\hat{\boldsymbol{\theta}}_i^{AUG}$ is a vector of Wainer's augmented subscore estimates for examinee *i*; $\overline{\hat{\boldsymbol{\theta}}^{*}}$ is the mean vector of unaugmented IRT subscore estimates; $\hat{\boldsymbol{\theta}}_i^{*}$ is a vector of unaugemented IRT subscore estimates for examinee *i*.

In order to evaluate the performance of Wainer's AUG, Wainer et. al (2001) applied their augmented subscoring method on three different types of observed subscores (classical standardized summed subscores, MAP subscores for response patterns, and EAP subscores for summed scores) in three operational tests (a certificate exam, a computer skill test, and an end-of-grade mathematics test). Among these three tests, the certificate exam appeared to be obviously unidimensional, whereas the computer skill test tended to be multidimensional. Regarding the end-of-grade mathematics test, the near-collinearity occurred.

Comparatively speaking, when what all subtests measured was homogenous, subscores estimated by Wainer's AUG were much more stable than the observed subscores. Nevertheless, these augmented subscores could not provide much diagnostic information distinguishable from the total score for individual examinees. Rather than basically replicating the total score as they do in unidimensional tests, Wainer's augmented subscores might be more discriminable in the case of multidimensional tests. That is, the largest weight in $\boldsymbol{B}$ or $\boldsymbol{B}^*$ is assigned to the target subtest and simultaneously the collateral information from the other subtests is borrowed to an extent depending on their reliabilities and intercorrelations manifested in $\boldsymbol{B}$ or $\boldsymbol{B}^*$. In the meantime, the stability of Wainer's augmented subscores remained substantially the same as it is in unidimensional tests.

Under some conditions, the stability could be breached, for which Wainer et. al (2001) suggested to increase the lengths of some or even all subtests. It was worth noting that adding more items in one subtest could also simultaneously enhance the reliabilities of subscores in other subtests. Moreover, Wainer's AUG treated the other subtests separately by assigning them different weights. In this aspect, it was more suitable and flexible than Yen's OPI when the test was essentially multidimensional. Yen's OPI procedure was principally established under the assumption of unidimensionality of the entire test. However, as a regression-based empirical Bayes subscoring approach, Wainer's AUG was vulnerable and dysfunctional when the collinearity occurred. The negative impacts were manifested as inconsistency of weights in $\boldsymbol{B}$ or $\boldsymbol{B}^*$ across test forms and aberrance of weights assigned to some observed target subscores. To reduce the existence of high correlations between some subtests, Wainer et. al (2001) reorganized the test by splitting or combining subtests.

In addition, a simulation study (Edwards & Vevea, 2006) was also conducted on the traditional summed scores and EAP summed scores under different simulated testing conditions. The number of subtests, subtest lengths, and correlations between subtests were manipulated in the study. Compared to the unaugmented subscores, Wainer's AUG procedure globally improved subscore estimates by means of yielding lower RMSE, higher reliability, and more accurate classifications. The similar result regarding improved reliability was also found in Skorupski and Carvajal's (2010) empirical study. In the meantime, Edwards and Vevea (2006) indicated that the magnitude of the improvement varied as the correlations between subtests and subtest lengths altered. Among a variety of simulated conditions, they stressed that the largest improvement was accomplished under conditions that the correlations between subtests were high, the reliability of observed target subscores was low, but the reliability of observed subscores in the other subtests was high.

As mentioned in Chapter 1, Luo, Diao, and Ren (2014) expanded Wainer's AUG to the CAT framework (AUG-CAT). According to their study, the augmentation techniques of Wainer's AUG were actually implemented after all the original MAP subscores were obtained. That is, prior to the augmentation procedure, the test was administered as a conventional CAT test consisting of a test battery, in which the examinees took the subtests one after the other in a fixed predetermined sequence. The administration of a CAT test battery in AUG-CAT exactly followed the same procedure as it is in the multiple independent UCAT (IND-UCAT). During the administration of each subtest, the MPI criterion and the MAP scoring algorithm (if the current study was considered) were employed for item selection and subscore estimation procedures. When an examinee completed a subtest, his/her MAP subscore for that subtest was correspondingly obtained. Then the examinee moved forward to the next subtest and the process described was repeated until the examinee finished all the subtests. Once the entire test was completed, all the MAP subscores of that examinee were obtained. Thereafter, the augmentation procedure regarding Wainer's AUG described previously was implemented on the MAP subscores estimated by IND-UCAT.

Subscoring by Adaptively Sequencing A Test Battery (SEQ-CAT) and reSEQ-CAT

Subscoring by adaptively sequencing a test battery (SEQ-CAT, see van der Linden, 2010) is an empirical Bayes subscoring approach, which primarily consists of a two-stage adaptive testing procedure in conjunction with the multilevel IRT modeling. It is one of the subscoring methods initially designed under the framework of computerized adaptive tests. For SEQ-CAT, the two-stage adaptive testing procedure indicates (1) the between-subtest adaptation determining the sequence of subtests administered to each examinee, and (2) the within-subtest adaptation determining the sequence of items administered to each examinee in a selected subtest.

The adaptation in the between-subtest stage reveals the principal difference between SEQ-CAT and IND-UCAT on the administration of a test battery, which may further enhance the testing efficiency of a CAT test. In IND-UCAT, the sequence of administering subtests is always predetermined and fixed to all examinees. By contrast, the optimal sequence of subtests is administered to individual examinees in SEQ-CAT. That is, each examinee may be provided with different orders of subtests according to their performance in the preceding subtests. The principle to optimize the sequence of subtests in SEQ-CAT is to screen each of the unadministered item subpools for the one with the largest sum of the prior expected Fisher's information across the intended subtest length. The prior expected Fisher's information is calculated by integrating the information function for each unadministered subtest over its own predictive posterior distribution, which is the updated joint marginal distribution by the responses from all the previous subtests.

The multilevel IRT modeling in SEQ-CAT refers to any applicable IRT models as first-level models and the specification of the joint distribution of all subscale parameters as a second-level model. The IRT models in the first level can be different, but not necessarily different for multiple unidimensional item subpools. The joint distribution in the second level contains the information on the associations between subscales, which is of great value to subscore estimation. Once the joint distribution of all subscale parameters $f(\theta_{(1)}, \theta_{(2)}, ..., \theta_{(D)})$ is specified, any marginal distribution or joint marginal distribution of the target subscales can be obtained by integrating the joint distribution of all subscales over all the other subscale dimensions. When each subtest is completed, the relevant joint marginal distribution is updated by the responses from all the preceding subtests, and is converted to the posterior distribution for the corresponding candidate subtest by integrating it over all the preceding subtest dimensions. The relevant joint marginal distribution refers to the joint distribution of all the administered subtests and any candidate subtest out of the unadministered subtests. The posterior distributions, also called the predictive posterior distributions mentioned above, are hereafter treated as prior distributions for selecting the next optimal subtest from all the unadministered subtests.

This procedure described above manifests the empirical Bayes algorithm in SEQ-CAT, which is the shifting process from the posterior distribution to the prior distribution by exploiting the collateral information obtained from the response vectors in all previous subtests. Once a subpool is identified, its posterior distribution is also used as the prior distribution for selecting the first item from that subpool. Then this prior distribution is continuously updated right after each selected item is completed by the examinee, so that the next optimal item could be selected until the prespecified subtest length or the accuracy criterion is reached. This item selection procedure used in SEQ-CAT reflects the MPI criterion, which is discussed in Section 2 of this chapter.

More precisely, as the second-level model in SEQ-CAT, the joint distribution of all subscale parameters should be specified in advance. For example,

$$f(\theta_{(1)}, \theta_{(2)}, ..., \theta_{(D)}) = \mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}), \tag{38}$$

by which the joint distribution is assumed to be a multivariate normal distribution, which is typically estimated from the field test. As described above, the selection of the optimal subtest over the other subtests is determined by the sum of the prior expected Fisher's information over the intended subtest length. As no subtest is administered yet, the respective marginal distribution $f(\theta_{(d)})$ of Equation (38) is used as the prior distribution for each subscale to compute the prior expected Fisher's information for each item from its own subpool, which is written as

$$\int I(\theta_{(d)}, u_{j(d)}) f(\theta_{(d)}) d\theta_{(d)} \,. \tag{39}$$

For comparing the sum of the prior expected Fisher's information across the intended subtest length among subpools, a shadow test is calculated for each subpool so that the items with the largest prior expected Fisher's information within each subpool can be selected, and also some constraints can be simultaneously satisfied. In the study, the length of each subtest is identical and fixed. Therefore, as the first subtest is to be selected, the objective function of the shadow test for subpool *d* is expressed as

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[\int I(\theta_{(d)}, u_{n_{(d)}}) f(\theta_{(d)}) d\theta_{(d)} \right] x_{n_{(d)}} \tag{40}$$

subject to

$$\sum_{n_{(d)}=1}^{N_{(d)}} x_{n_{(d)}} = J_{(d)} \tag{41}$$

$$x_{n_{(d)}} \in \{0,1\},\ n_{(d)} = 1, 2, ..., N_{(d)} \,. \tag{42}$$

By Equations (40) to (42), the first subpool d^1 is identified and then the item that has the maximum prior expected information in the shadow test is administered as the first item from subpool d^1. Its response is correspondingly utilized to update the prior distribution $f(\theta_{(d^1)})$ by Equation (6) and the posterior distribution $f(\theta_{(d^1)} \mid u_{i(d^1)})$ is thereafter obtained. Then the posterior distribution is to fit into Equation (23) to calculate the second shadow test for subscale d^1, and the item with the maximum information is again selected to be administered, whose response is also used to update the posterior distribution. The process continues as described in the sections of MPI criterion and shadow test for UCAT of this chapter. For legible guidance, Equations (23) to (26) for selecting the *k*th item from subpool d^1 is generalized as

$$\text{maximize} \sum_{n_{(d^1)}=1}^{N_{(d^1)}} \left[\int I(\theta_{(d^1)}, u_{n_{(d^1)}}) f(\theta_{(d^1)} \mid \boldsymbol{u}_{i(d^1)}^{k-1}) d\theta_{(d^1)} \right] x_{n_{(d^1)}} \tag{43}$$

subject to

$$\sum_{n_{(d^1)}=1}^{N_{(d^1)}} x_{n_{(d^1)}} = J_{(d^1)} \tag{44}$$

$$x_{n_{(d^1)}} = 1, \textit{ for all } n_{(d^1)} \in S_{(k-1)_{(d^1)}} \tag{45}$$

$$x_{n_{(d^1)}} \in \{0,1\},\ n_{(d^1)} = 1, 2, ..., N_{(d^1)}. \tag{46}$$

When the subtest d^1 reaches the predetermined length, the final updated posterior distribution is used to estimate the MAP score of subtest d^1, which is the solution to Equation (47),

$$\frac{\partial}{\partial\theta_{(d^1)}} \ln f(\theta_{(d^1)} \mid \boldsymbol{u}_{i(d^1)}) = 0\,. \tag{47}$$

As the second subtest is to be selected, the prior distribution $f(\theta_{(d)})$ in Equation (40) is updated as the posterior distribution $f(\theta_{(d)} \mid \boldsymbol{u}_{i(d^1)})$ $(d \neq d^1)$, which is the relevant joint marginal distribution for any candidate subtest updated by the responses from the first administered subtest and integrated over the same subtest dimension (see the first step in Equation (48)). That is, there are still *D*-1 unadministered subtests and therefore altogether *D*-1 posterior distributions would be calculated. For the sake of clarification, the posterior distributions used for selecting the next optimal subtest are deducted from

$$\begin{aligned} f(\theta_{(d)} \mid \boldsymbol{u}_{i(d^1)}) &= \int f(\theta_{(d)}, \theta_{(d^1)} \mid \boldsymbol{u}_{i(d^1)}) d\theta_{(d^1)} \\ &= \int f(\theta_{(d)} | \theta_{(d^1)}) f(\theta_{(d^1)} \mid \boldsymbol{u}_{i(d^1)}) d\theta_{(d^1)} \\ &\propto \int f(\theta_{(d)} | \theta_{(d^1)}) f(\theta_{(d^1)}) L(\boldsymbol{u}_{i(d^1)} \mid \theta_{(d^1)}) d\theta_{(d^1)} \\ &= \int f(\theta_{(d)}, \theta_{(d^1)}) L(\boldsymbol{u}_{i(d^1)} \mid \theta_{(d^1)}) d\theta_{(d^1)} \end{aligned}, \qquad (48)$$

where $d \neq d^1$ (van der Linden, 2010). The second step in Equation (48) reflects the assumption of conditional independence of $\theta_{(d)}$ and $\boldsymbol{u}_{i(d^1)}$ given $\theta_{(d^1)}$, and the fact that $f(\theta_{(d)} \mid \boldsymbol{u}_{i(d^1)})$ is actually the predictive posterior distribution by marginalizing $f(\theta_{(d)} | \theta_{(d^1)})$ over the posterior distribution of $\theta_{(d^1)}$ given $\boldsymbol{u}_{i(d^1)}$. In the meantime, the last step provides a clue to compute $f(\theta_{(d)} \mid \boldsymbol{u}_{i(d^1)})$ in a more straightforward manner. By compared $D-1$ shadow tests obtained from Equations (40) to (42), the second subpool d^2 can be identified. Likewise, the posterior distributions fitting to Equation (40) for selecting the third subtest follows the same deduction and are denoted as

$$f(\theta_{(d)} \mid \boldsymbol{u}_{i(d^1)}, \boldsymbol{u}_{i(d^2)}) \propto \iint f(\theta_{(d)}, \theta_{(d^1)}, \theta_{(d^2)}) L(\boldsymbol{u}_{i(d^1)} \mid \theta_{(d^1)}) L(\boldsymbol{u}_{i(d^2)} \mid \theta_{(d^2)}) d\theta_{(d^1)} d\theta_{(d^2)}, \quad (49)$$

where $d \neq d^1 \neq d^2$.

As for the administration of the selected subtests, the process holds the same as implemented within the first selected subtest, except for substituting the posterior distribution $f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)})$ or $f(\theta_{(d^3)} \mid \boldsymbol{u}_{i(d^1)}, \boldsymbol{u}_{i(d^2)})$ for the prior distribution $f(\theta_{(d^1)})$ by Equation (6), after the first response in subtest d^2 or d^3 is obtained. As such, the following responses continuously update that posterior distribution for the next optimal item selection. For example, $f(\theta_{(d^2)} \mid \boldsymbol{u}^{k-1}_{i(d^2)}, \boldsymbol{u}_{i(d^1)})$ and $f(\theta_{(d^3)} \mid \boldsymbol{u}^{k-1}_{i(d^3)}, \boldsymbol{u}_{i(d^2)}, \boldsymbol{u}_{i(d^1)})$ are respectively the posterior distributions updated by the (*k*-1)th selected item in subtests d^2 and d^3. These posterior distributions are then fitted into Equations (43) to (46) to select the *k*th item for subtests d^2 and d^3.

To evaluate the efficiency of SEQ-CAT, a simulation study was conducted under the conditions of different subtest lengths and content constraint impositions, in contrast to the baseline method of IND-UCAT (van der Linden, 2010). The results indicated that the adaptive subtest sequencing could improve the accuracy of subscore estimates, even for the short test, by comparison to the baseline method. Also, the information borrowed from the earlier subtests was greatly beneficial to the ability estimation in the later subtests for the examinees at the two extreme ends of the ability scale. In addition, the constraints did not have a strong impact on the subscore estimation when the shadow test was conducted to impose the constraints in the study.

On the other hand, a concern may arise in SEQ-CAT. That is, the later the subtests are selected and administered, the more information they may take advantage of for subscore estimation. It is because more responses and the relevant joint marginal distribution involving more subscales are included for subscore estimation as the test proceeds, which is especially true compared to the very early administered subtests. With respect to this issue, van der Linden further developed SEQ-CAT and raised the post-hoc fully Bayesian subscore estimation of reSEQ-CAT (W. J. van der Linden, personal communication, July 30th, 2013). That is, when all the subtests are completed by an examinee, the subscores estimated by incomplete response patterns and the relevant joint marginal distribution are reestimated by reformulating their prior distribution. These subscores refer to the subscores in all subtests except the last subtest.

Specifically, the prior distribution of subtest *d* is reformulated with the joint distribution of all subscales updated by the responses from all the other subtests and then integrated over these subscales, as is conducted in Equations (48) and (49). The MAP score of subtest *d* is then reestimated by the solution to Equation (50),

$$\frac{\partial}{\partial \theta_{(d)}} \ln f(\theta_{(d)} \mid \boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(D)}) = 0\,, \tag{50}$$

where $f(\theta_{(d)} \mid \boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(D)})$ is the posterior distribution for subscale *d*, which is derived from the reformulated prior distribution and all the responses in subtest *d*. Since the last administered subtest has utilized the information provided by all the responses and the joint distribution of all subscales, its subscores do not require to be reestimated in reSEQ-CAT. Likewise, the EAP score of subtest *d* can also be obtained by the following equations:

$$\begin{gathered}\hat{\theta}_{(1)} = \int \theta_{(1)} f(\theta_{(1)} \mid \boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(D)}) d\theta_{(1)} \\ \vdots \\ \hat{\theta}_{(D-1)} = \int \theta_{(D-1)} f(\theta_{(D-1)} \mid \boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(D)}) d\theta_{(D-1)}\end{gathered} \quad . \tag{51}$$

As in AUG-CAT, reSEQ-CAT is also applicable in both P&P tests and CATs. Following the instructions from van der Linden, Liu, Li, and Choi (2014) applied reSEQ-CAT in their study by contrast to SEQ-CAT. The findings from their study showed that reSEQ-CAT could produce more accurate subscore estimates compared to SEQ-CAT as the correlations among subscales were medium or high. In the meantime, the total scores were calculated based on the reSEQ-CAT subscore estimates in their study. Those total scores were also recovered to a greater extent.

Subscoring by Multidimensional Adaptive Testing (MCAT)

Multidimensional adaptive testing (MCAT, see Segall, 1996) is an adaptation of the conventional multidimensional IRT (MIRT) subscoring method in the CAT testing environments. As in MIRT Bayesian scoring procedure, MCAT enhances the measurement efficiency simultaneously on multiple subscales by adding the information on the correlations among subscales to the score estimation procedure, in contrast to IND-UCAT that ignores the unique source of information underlying these subscales. Also, compared to the MIRT Bayesian scoring procedure, given the characteristics of adaptive tests, the measurement efficiency in MCAT intends to be further enhanced by customizing a real-time test corresponding to the performance of an examinee on all previous items. The improvement on score estimation is also validated even when MCAT demonstrates a simple structure unless the prior joint distribution of subscales is a diagonal matrix, for which MCAT is reduced and equivalent to IND-UCAT.

Regarding the scoring procedure in MCAT, the MAP subscore estimates have been discussed in the section of MAP estimates of this chapter. Another absolutely necessary procedure in MCAT is to specify the item selection criterion, which in the literature mainly includes maximizing the determinant of Fisher's information matrix or Fisher's posterior information matrix (D-optimality or a Bayesian version of D-optimality, see Luecht, 1996; Segall, 1996), minimizing the trace of the inverse of Fisher information matrix (A-optimality, see van der Linden, 1999), and maximizing the posterior expected Kullback-Leibler information (KLI, see Chang & Ying, 1996; Veldkamp & van der Linden, 2002). In line with the counterpart adopted in UCAT, the focus of the MCAT item selection criterion in the study lays on the Bayesian version of D-optimality, which possesses widespread recognition in the literature (Luecth, 1996; Li & Schafer, 2005; Wang & Chen, 2004; Lee, Ip, & Fuh, 2008; Allen, Ni, & Haley, 2008; Mulder & van der Linden, 2009).

Conceptually speaking, D-optimality is a criterion of selecting an item that most largely reduces the volume of the credibility ellipsoid from the rest of an item pool *R*. For a multivariate normal distribution, the volume of the credibility ellipsoid after administering the *k*th item is defined as

$$\varsigma \times |\mathbf{\Sigma}^{k}|^{1/2}, \tag{52}$$

where

$$\varsigma = \frac{2\pi^{D/2}[\chi_D^2(p)]^{D/2}}{D\Gamma(\frac{1}{2}D)} \tag{53}$$

and $\mathbf{\Sigma}^k$ is the covariance matrix calculated after the *k*th item is administered. In Equation (53), $\Gamma(\cdot)$ is the gamma function and $\chi_D^2(p)$ is the quantile function of a chi-squared distribution, χ_D^2, with *D* degrees of freedom for probability p. In other words, $\chi_D^2(p)$ is the value of χ_D^2 at the $p \times 100$ percentile. Equation (53) shows that ς is merely a function of *D* and p, so it always holds constant across items in a test. Therefore, the decrement V_k on the volume of the credibility ellipsoid only depends on the decrement from $\mathbf{\Sigma}^{k-1}$ to $\mathbf{\Sigma}^k$. To be more explicit,

$$\begin{aligned} V_k &= \varsigma\,|\mathbf{\Sigma}^{k-1}|^{1/2} - \varsigma\,|\mathbf{\Sigma}^k|^{1/2} \\ &= \varsigma(|\mathbf{\Sigma}^{k-1}|^{1/2} - |\mathbf{\Sigma}^k|^{1/2}) \end{aligned}. \tag{54}$$

For the IRT maximum likelihood estimates, the covariance matrix could be approximated by the inverse of Fisher's information matrix $\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$. Also, the determinant of the inverse of a matrix is algebraically equal to the reciprocal of the determinant. Therefore, Equation (54) can be rewritten as

$$\begin{aligned} V_k &= \varsigma\,|\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k-1})^{-1}|^{1/2} - \varsigma\,|\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k})^{-1}|^{1/2} \\ &= \varsigma\,|\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k-1})|^{-(1/2)} - \varsigma\,|\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k})|^{-(1/2)} \end{aligned}, \tag{55}$$

where $\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k-1})=\sum_{j=1}^{k-1}\boldsymbol{I}(\hat{\boldsymbol{\theta}}^{k-1},u_j)$. Since the information function is additive within a test,

$$\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k})=\boldsymbol{I}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{k-1})+\boldsymbol{I}(\hat{\boldsymbol{\theta}}^{k-1},u_k)\,. \tag{56}$$

In Equations (55) and (56), the first term on the right hand side is all constant for all the remaining items in the pool R_k and therefore the magnitude of V_k is determined only by $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_i^{k-1},u_k)$. Apparently, V_k can be maximized by an item that maximizes the determinant of Equation (56), that is,

$$\arg\max_{k\in R_k}\{\det(\boldsymbol{I}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1})+\boldsymbol{I}(\hat{\boldsymbol{\theta}}_i^{k-1},u_k))\}\,. \tag{57}$$

Furthermore, MAP estimates obtained from a multivariate posterior distribution correspond to the item selection criterion of the Bayesian version of D-optimality, for which Equations (55) to (57) are also applicable. For ease of exposition, Equation (55) for a multivariate posterior distribution is rewritten as

$$\begin{aligned} C_k &= \varsigma\,|\boldsymbol{W}_i^{k-1}|^{1/2}-\varsigma\,|\boldsymbol{W}_i^{k}|^{1/2} \\ &= \varsigma(|\boldsymbol{W}_i^{k-1}|^{1/2}-|\boldsymbol{W}_i^{k}|^{1/2}) \end{aligned}, \tag{58}$$

where C_k is the decrement on the volume of the posterior credibility ellipsoid and $\boldsymbol{W}_i^{k-1}$ is the posterior covariance matrix calculated after the (*k*-1)th item is administered. Similarly, the posterior covariance matrix $\boldsymbol{W}_i^{k-1}$ can be approximated by the inverse of Fisher's posterior information matrix $\boldsymbol{I}^{p}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1})$, which is given by

$$\boldsymbol{I}^{p}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) = \boldsymbol{I}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) + \boldsymbol{\Phi}^{-1}, \tag{59}$$

where $\boldsymbol{\Phi}$ is the prior covariance matrix. According to Equations (55) and (56), $\boldsymbol{W}_i^{k-1}$ in Equation (58) is substituted by Equation (59). Therefore,

$$\begin{aligned} C_k &= \varsigma \,|\, \boldsymbol{I}^{p}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) \,|^{-(1/2)} - \varsigma \,|\, \boldsymbol{I}^{p}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}_i^{k-1},u_k) \,|^{-(1/2)} \\ &= \varsigma \,|\, \boldsymbol{I}^{p}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) \,|^{-(1/2)} - \varsigma \,|\, \boldsymbol{I}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}_i^{k-1},u_k) + \boldsymbol{\Phi}^{-1} \,|^{-(1/2)} \end{aligned} . \tag{60}$$

Because the first term in Equation (60) is constant for all the remaining items in the pool R_k, C_k is maximized by selecting an item that could maximize the second determinant in the equation, that is,

$$\arg\max_{k\in R_k}\{\det(\boldsymbol{I}(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i^{k-1}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}_i^{k-1},u_k) + \boldsymbol{\Phi}^{-1})\} . \tag{61}$$

By comparing Equations (61) and (57), it shows that the Bayesian version of D-optimality is expected to outperform the conventional D-optimality, because more information from the prior distribution is provided and utilized for item selection in MCAT. However, this finding is not applicable to a simple-structure MCAT test, which is further explained in the fourth section of Chapter 5. On the other hand, as indicated by Segall (1996), the item selection under the D-optimality may lead to indefinable or poorly definable ability estimates, which is especially validated in a simple-structure test battery. Given the feature of simple structure that each item loads only on one subscale, when the first item is to be selected by D-optimality criterion, the information matrix produced by any individual item out of the pool *R* (Equation (56)) is a diagonal matrix with only one diagonal element nonzero. It means that the elements in the other two rows are all zeros. Theoretically, the determinant of such an information matrix is always zero, which implies that the determinants of the information matrices yielded by all the individual items in the pool are all zero. This non-distinction of determinants among all the items results in the deadlock of the item selection procedure, and therefore the provisional ability estimates are indefinable. Under such a circumstance, the efficiency of MCAT will be appallingly impeded, unless the first three items (if the current study is considered) are enforced to be selected from three different subscales or the simplified item selection criterion is applied, which is described in the fourth section of Chapter 5.

Generally speaking, by contrast with UCAT, the increase of measurement efficiency in MCAT is primarily manifested as greater precision, shorter test length and higher score reliability. As demonstrated in Segall’s (1996) study, MCAT could provide equal or higher precision by saving about one-third items compared to UCAT when moderate to high correlations existed among subtests. Also, MCAT achieved considerably greater improvement in reliability when holding the same test length as UCAT. The findings were also aligned with the results of Luecht’s (1996) study in which complex content constraints were imposed.

For some circumstances, imposing content constraints facilitates implanting subscoring mechanism in MCAT, especially when the discriminating power and difficulty levels of items among subpools are not balanced. First, content constraints ensure that examinees could reach the items from all content areas that are required in test specifications. Second, without much compromise of adding items that may be resourceful to some subtest, but redundant or non-informative to the other subtests as it does in UIRT and UCAT, an optimal item from any subtest in MCAT could provide information for estimating and updating all subscores to some extent, depending on the dependencies among subscales or content areas unless the dependencies never exist. Third, content constraints can reduce a large discrepancy on the number of items from each content area, which may result from the confounding between contents and item difficulties. In practice, according to the test specifications, some other constraints can also be imposed such as word counts, item exposure control, and test length.

Li and Schafer (2005) applied MCAT to a test battery involving Reading and Math in a real testing program. They pointed out that compared to IND-UCAT, MCAT can increase the rate of item utilization in the pool and yield more accurate subscore estimates, even for the examinees at the extreme ability levels. Wang and Chen (2004) conducted a simulation study to investigate the measurement efficiency of MCAT on polytomous items and complex-structure items. They concluded that MCAT performed more efficiently than UCAT and IND-UCAT as the correlations between subtests, the number of subtests, and the number of scoring levels increased. As a matter of fact, comparing the five subscoring methods in this study essentially reflects how UCAT and MCAT differ in utilizing the collateral information in subscore estimation. In principle, IND-UCAT, AUG-CAT, SEQ-CAT, and reSEQ-CAT are actually the different manifestations of UCATs, and PC-MCAT is a modified MCAT. In addition, among these four types of UCATs, different fashions of adding collateral information to subscore estimation are also demonstrated in the study.

Higher-Order IRT Model (HO-IRT)

The one-factor higher-order IRT model (HO-IRT, see de la Torre & Song, 2009) is usually adopted for modeling the item response data with a hierarchical latent trait structure in modern assessments. As discussed in Chapter 1, the hierarchical latent trait structure exhibits a two-order structure with multiple subscale abilities as the first order and a general ability as the second order. The multiple subscale abilities are, in general, measured by a test battery, which contains multiple subtests, each targeted at some specific content or skill. For such response data, the use of a unidimensional IRT model on the entire test will violate the assumption of unidimensionality. Conducting multiple independent UIRT models (IND-UIRT) will ignore the associations among subscales. Considering a MIRT model will overlook the hierarchical structure between subscale abilities and the general ability. As the most complex model among the models mentioned above, the HO-IRT model can fairly account for the multidimensionality, the correlations among subscales, and the hierarchy between different levels of latent traits as a whole in one model.

Also, the HO-IRT model allows for the applicability of multiple identical or different conventional unidimensional IRT models to the multidimensional response data, which are directly dominated by multiple first-order abilities. This application procedure of estimating the first-order abilities is conducted in the measurement phase of the HO-IRT model. In the meantime, a large amount of variance shared by the first-order subscale abilities is accounted for by the second-order general ability, which can also be estimable at the structure phase of the HO-IRT model. The nature of integrating two levels of latent traits in one model determines that the parameters for the first-order and second-order latent traits can be estimated simultaneously in the HO-IRT model. These parameters are then converted to some certain scale scores, which are operationally reported as subscores and the total score respectively.

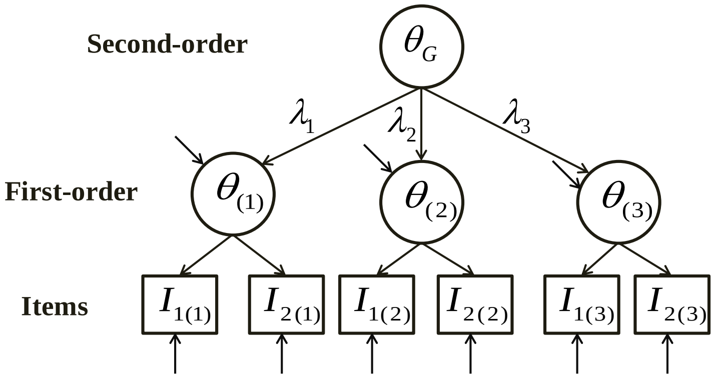

Figure 1. Example of the One-Factor Higher-Order IRT Model.

Figure 1 above presents an example of the one-factor HO-IRT model. In the model, the second-order latent trait, which is referred to as the general ability θ_{iG} in the study, typically follows a standard normal distribution $\theta_{iG} \sim N(0,1)$ in educational assessments. It demonstrates the modeling of the joint distribution of the first-order latent traits, which are referred to as subscale abilities $\theta_{i(d)}$ in the study. The loading $\lambda_{(d)}$ of subscale *d* on the general ability reflects its correlation to the general ability and is also viewed as the regression coefficient of subscale *d* on the general ability in the linear function. That is, each subscale ability in HO-IRT model is linearly correlated to the general ability and can be expressed by a linear function of the general ability

$$\theta_{i(d)} = \lambda_d \theta_{iG} + \varepsilon_{i(d)}\ , \qquad (62)$$

where $|\lambda_d| \leq 1$ and is always expected to be nonnegative in most cases due to the nature of the relationships among abilities in reality. Also, $\varepsilon_{i(d)}$ is the disturbance for subscale *d* with the distribution of $\varepsilon_{i(d)} \sim N(0,1-\lambda_d^2)$ and is independent of other disturbances and all abilities.

Note that the constraints on the magnitude of λ_d are very necessary in the HO-IRT model because they can make the general ability and all the subscale abilities estimated on the same scale. In other words, under the constraints, the marginal distribution of each subscale ability all follows the standard normal distribution, $\theta_{(d)} \sim N(0,1)$, as with the distribution of the general ability. In addition, the product of two loadings reflects the correlation between two subscale abilities. Therefore, the correlation matrix of the subscale abilities in the model of Figure 1 is shown as

$$\begin{array}{c} \\ \theta_{(1)} \\ \theta_{(2)} \\ \theta_{(3)} \end{array} \begin{array}{c} \begin{array}{ccc} \theta_{(1)} & \theta_{(2)} & \theta_{(3)} \end{array} \\ \begin{bmatrix} 1 & & \\ \lambda_1\lambda_2 & 1 & \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & 1 \end{bmatrix} \end{array} . \qquad (63)$$

Furthermore, conditional on the general ability θ_{iG}, the subscales are independent of each other and each follows the distribution of $\theta_{i(d)} | \theta_{iG}, \lambda_d \sim N(\lambda_d \theta_{iG}, 1-\lambda_d^2)$. Because the unidimensionality exists in each subtest due to the simple structure, the conventional UIRT model (i.e. 1PL, 2PL, or 3PL UIRT model) can be applied in each subtest.

Given the structure of the HO-IRT model, all the unknown parameters of interest can be estimated jointly including item parameters, the general ability parameter, subscale ability parameters, and the regression coefficients (the loadings). However, due to its complexity and high ability dimensionality involved, the joint estimation of all the parameters have to be conducted by using MCMC algorithm based on the hierarchical Bayesian formulation (Sheng & Wikle, 2007; de la Torre & Hong, 2010; Huang, Wang, Chen, & Su, 2013). Although many studies have provided strong supports on the HO-IRT model regarding their accuracy on parameter estimates compared to IND-UIRT (de la Torre & Song, 2009; de la Torre & Hong, 2010; de la Torre, Song, & Hong, 2011), the demands of intensive computations on MCMC estimation for the HO-IRT model largely confine its application in practice, especially their use in the routine scoring procedure. By taking advantage of the hierarchical latent trait structure in the HO-IRT model, the current study suggests a successive scoring procedure to calculate the total scores based on the subscores estimated by the five compared subscoring methods. This suggested procedure requires much less computation by assuming that all the loadings are known, which is described in detail in Section 3 of Chapter 3.

Primary Comparison Studies on Some Subscoring Methods

To better evaluate the performance of the existing subscoring methods, substantial comparison studies were conducted in the context of both simulated data and empirical data in the recent decade. The findings from these studies provided some constructive and practical guidelines for future research and the operational uses of these methods. Tables 1 and 2 in the first chapter enumerated the primary comparison studies related to the subscoring methods discussed above, and all of these studies apparently focused only on the P&P tests. Despite the discrepancy in testing formats, the studies can still provide very instructive perspectives on the implementation of the comparison study in the CAT framework.

Some of the comparison studies, partly listed in Tables 1 and 2, employed the multiple independent unidimensional IRT model (IND-UIRT) as the baseline method, so as to demonstrate the effects of utilizing the collateral information among subscales on subscore estimation (DeMars, 2005; Yao, 2010; van der Linden, 2010). Some other studies adopted the proportion-correct (PC) subscoring as the baseline method, which is virtually a classical version of IND-UIRT (Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Shin, 2007; Yao & Boughton, 2007). There were also a few studies using both as the baseline method (Edwards & Vevea, 2006; Fu & Qu, 2010). Numerous previous studies have suggested that the classical scoring method has many limitations on estimating ability parameters compared to the IRT scale scoring such as low reliability and sample dependence. Consequently, use of the PC subscoring as the baseline method may introduce some more disturbances to the comparison, such as the differences of the augmented IRT subscores (i.e. MAP or EAP estimates) versus the classical unaugmented PC subscores.

Furthermore, all of the comparison studies adopted either a simulation design or empirical data or both as listed in Table 2. For simulation studies, if the MIRT or HO-IRT model were involved in the comparison, either the 3-parameter logistic (3PL) MIRT model or the higher-order IRT model or both were employed to generate the response patterns (DeMars, 2005; Yao, 2010; Yao & Boughton, 2007; de la Torre, Song, & Hong, 2011). Otherwise, the unidimensional IRT model was used, which was always the 3PL model (Edwards & Vevea, 2006; Shin, 2007; van der Linden, 2010). As a matter of fact, the 3PL MIRT or UIRT model was also largely used in empirical studies to estimate subscores (DeMars, 2005; Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Stone, Ye, Zhu, & Lane, 2010; Skorupski & Carvaljal, 2010). The item parameters adopted in the simulations studies were typically drawn from the real item pool (DeMars, 2005; Shin, 2007; van der Linden, 2010; Yao, 2010; de la Torre, Song, & Hong, 2011) and only a few studies simulated item parameters for their own use (Edwards & Vevea, 2006; Fu & Qu, 2010). For the former case, the item parameters were assumed to be known so that the studies were dedicated to the subscoring procedure. There were also a number of studies that estimated both item parameters and subscores after they simulated the responses. The reason for doing it was because they intended to take the errors from both estimations into account (DeMars, 2005; Edwards & Vevea, 2006; Yao & Boughton, 2006; Fu & Qu, 2010; Yao, 2010).

Regarding the empirical studies, both unidimensional and multidimensional real data were investigated for comparisons of the subscoring methods (Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Stone, Ye, Zhu, & Lane, 2010). Also, a simple structure of items in MIRT was assumed across the empirical and simulation studies when the MIRT subscoring method was compared to other methods (Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Yao & Boughton, 2007; Stone, Ye, Zhu, & Lane, 2010; Fu & Qu, 2010; Yao, 2010; de la Torre, Song, & Hong, 2011). Moreover, a variety of studies evaluated the subscoring methods on dichotomous items (DeMars, 2005; Edwards & Vevea, 2006; Skorupski & Caravajal, 2010; de la Torre, Song, & Hong, 2011) and some were conducted on the mixed item type with both dichotomous and polytomous items included in each subtest (Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Shin, 2007; Yao & Boughton, 2007; Stone, Ye, Zhu, & Lane, 2010; Yao, 2010). Root mean square error (RMSE) and bias (absolute bias, bias, or conditional bias) were the most commonly-used outcome measures to evaluate the recovery of subscale scores in almost all of the simulation studies listed in Tables 1 and 2. Empirical studies typically adopted the descriptive statistics like the mean and the standard deviation (SD) of the subscore estimates as outcome measures (DeMars, 2005; Stone, Ye, Zhu, & Lane, 2010; de la Torre, Song, & Hong, 2011). Some of the studies also employed the reliability or/and the correlations between true subscores and estimated subscores as additional outcome measures (DeMars, 2005; Edwards & Vevea, 2006; Shin, 2007; Skorupski & Caravajal, 2010).

DeMars (2005) compared three subscoring models (bifactor models, MIRT models, and Wainer's AUG) to the IND-UIRT model and indicated that compared to the IND-UIRT model, the three subscoring methods produced comparably low bias and RMSE when the highly-correlated subscale abilities were measured by a test with moderately short subtest length (15-20 items) and when these subtests were administered at one time. Some other studies also reached the similar conclusion when they compared different subscoring methods (MIRT, HO-IRT, SEQ-CAT, and AUG) to the baseline method (either the IND-UIRT model or/and the PC method) (Dwyer, Boughton, Yao, Lewis, &

Steffen, 2006; Shin, 2007; Fu & Qu, 2010;Yao, 2010; van der Linden, 2010). These findings strongly suggested that the proper use of the collateral information could improve or at least not impair the subscore estimates.

DeMars (2005) also pointed out that MIRT and AUG produced relatively smaller standard errors and less bias on estimated subscores, which was aligned with the results regarding MIRT models in Yao's (2010) study, in which a test with longer subtest length (34 to 57 items) measured multiple subscales with low or zero correlations by mixed item types. Yao (2010) also found that HO-IRT models and MIRT models performed comparably well on the recovery of subscores and total scores compared to bifactor models. As a matter of fact, when AUG, MIRT, and HO-IRT were selected for comparison, substantial studies suggested that they always outperformed over the other methods (e.g. OPI and bifactor models) on subscore estimation and that the differences between these three methods were minor (Dwyer, Boughton, Yao, Lewis, & Steffen, 2006; Shin, 2007; Fu & Qu, 2010; Yao, 2010; de la Torre, Song, & Hong, 2011). For examinees with extreme abilities, MIRT and HO-IRT performed even better (de la Torre, Song, & Hong, 2011). However, AUG may attract more favorable attention in practice due to its relatively unsophisticated computations. In addition, OPI could actually perform comparably to AUG and MIRT on subscore estimation when the correlations between subscale abilities were high, whereas OPI might also produce even larger RMSE than both of the baseline methods (IND-UCAT and PC subscoring) as the correlations were low (Yao & Boughton, 2007; Fu & Qu, 2010; de la Torre, Song, & Hong, 2011).

Regarding the factors affecting subscore estimation, Shin (2007) indicated that the subtest length and the correlations between subscales could affect the magnitude of RMSE, SD, bias, and reliability. High correlations between subscales and the increase of subtest length would improve the accuracy of ability parameter estimation for the subscoring methods, especially for OPI (Fu & Qu, 2010). In addition, an increase of subtest length could, to some extent, offset the negative impact imposed from low correlations between subscales (Yao, 2010). The sample size is not an influential factor in subscore estimation if item parameters are assumed to be known in the studies, but it is crucial to item parameter estimation in the item calibration process (Shin, 2007; Yao & Boughton, 2007; de la Torre, Song, & Hong, 2011). de la Torre, Song, and Hong (2011) implemented the only study that investigated the effect of the number of subtests on subscore estimates among four subscoring methods (MIRT, HO-IRT, AUG, and OPI). They pointed out that more subtests in a test battery could improve the correlations between true subscores and estimated subscores, but did not demonstrate a noticeable impact on reducing RMSE. Despite the same factor also investigated by Edwards and Vevea (2006), their study placed more focus on the comparisons of different subscore types (summed scores versus the IRT scale scores for summed scores) for Wainer's AUG, instead of the comparisons of different subscoring methods.

CHAPTER III

RESEARCH METHODOLOGY

With reference to the previous comparison studies on subscoring methods in P&P tests, a simulation study was designed and conducted to evaluate the performance of five subscoring methods in CAT, of which the conventional MCAT was modified as the pool-constrained MCAT (PC-MCAT) for more realistic comparisons. A variety of testing conditions, depending on the combinations of different levels of two primary factors, were also simulated for the purpose of investigating their effects on subscore estimation. In the meantime, each component consisting of a CAT test was specified in detail and remained consistent across all the five compared methods. Finally, the suggested total score estimation approach was illustrated as the second stage of the successive scoring procedure proposed in the study.

Simulation Design

In the study, five different CAT subscoring methods, AUG-CAT, SEQ-CAT, reSEQ-CAT, PC-MCAT, and IND-UCAT, were examined, of which IND-UCAT was adopted as the baseline method. PC-MCAT is a modified MCAT, which is further described in the next section of this chapter. Also, the subscoring procedures implemented in AUG-CAT and reSEQ-CAT are actually the same as their applications in P&P tests. These two methods were suffixed with “CAT” merely because they were applied in the CAT tests in the study. As the post-hoc subscoring methods, their estimation algorithms are only conducted after an entire test is completed, which implies that they do not intervene in any stage of the CAT testing process such as the item selection and MAP subscoring procedures. On the other hand, because IND-UCAT, SEQ-CAT, and PC-MCAT are initially designed under the CAT framework, they possess their own algorithm for item selection while conducting the just-in-time subscoring procedure. Given that the intent of the study concentrated on the comparisons of different subscoring procedures among the five methods, the differences on the item selection procedures should be ruled out for more convincing comparisons. As a consequence, the five subscoring methods were respectively paired with each of the three item selection algorithms. That is, three sets of items selected by IND-UCAT, SEQ-CAT, and PC-MCAT were individually scored by each of the five subscoring methods. Because the sample size is not influential to the subscore estimation in this study design, only one sample size ($I = 1,000$) was considered. In addition, a three-subtest ($D = 3$) test battery was investigated in the study.

Table 3

Loadings in Three Correlation Structures

Correlation	$\lambda_{(1)}$	$\lambda_{(2)}$	$\lambda_{(3)}$
Low	.45	.50	.55
Mixed	.50	.95	.80
High	.93	.95	.98

As mentioned previously, two primary factors were varied in the study: subtest length (J = 10 and 20) and the correlations between subtests. In terms of the last factor, three different correlation structures (low, mixed, and high) were considered. In HO-IRT, the magnitudes of the correlations $\rho_{(dd')}$ among subtests are determined by the loadings $\lambda_{(d)}$ of these subtests on the general ability. The specific values of the loadings shown in Table 3 were all arbitrarily assumed in the study. According to Equation (63), the corresponding correlation matrices for the levels of low, mixed, and high are respectively expressed as

$$\begin{array}{cc} & \begin{array}{ccc} \theta_{(1)} & \theta_{(2)} & \theta_{(3)} \end{array} \\ \begin{array}{c} \theta_{(1)} \\ \theta_{(2)} \\ \theta_{(3)} \end{array} & \left[\begin{array}{ccc} 1 & & \\ .23 & 1 & \\ .25 & .28 & 1 \end{array}\right] \end{array}, \quad \begin{array}{cc} & \begin{array}{ccc} \theta_{(1)} & \theta_{(2)} & \theta_{(3)} \end{array} \\ \begin{array}{c} \theta_{(1)} \\ \theta_{(2)} \\ \theta_{(3)} \end{array} & \left[\begin{array}{ccc} 1 & & \\ .48 & 1 & \\ .40 & .76 & 1 \end{array}\right] \end{array}, \quad \begin{array}{cc} & \begin{array}{ccc} \theta_{(1)} & \theta_{(2)} & \theta_{(3)} \end{array} \\ \begin{array}{c} \theta_{(1)} \\ \theta_{(2)} \\ \theta_{(3)} \end{array} & \left[\begin{array}{ccc} 1 & & \\ .88 & 1 & \\ .91 & .93 & 1 \end{array}\right] \end{array}. \qquad (64)$$

In the study, the HO-IRT model was employed to generate different orders of the true ability parameters (subscale ability parameters and the general ability parameters). In addition, the hierarchical structure of abilities in HO-IRT provides the possibility of conducting the successive scoring procedure proposed in the study, which is elaborated in Section 3 of this chapter. Assume that the distribution of the general ability for the population of examinees was a standard normal distribution. 1,000 examinees with different levels of the general ability were randomly drawn from $\theta_G \sim N(0,1)$. As described in Section 7 of Chapter 2, given the general ability θ_{iG} of examinee *i*, his/her subscale parameters were correspondingly generated from $\theta_{i(d)} \mid \theta_{iG}, \lambda \sim N(\lambda\theta_{iG}, 1-\lambda^2)$. Table 4 and Figure 2 below present the descriptive summary and the distributions of different orders of ability parameters simulated in the study, which were considerably aligned with the simulation design. Because the scoring procedure is the primary interest in the study, item parameters and the loadings were assumed to be known, which was also in line with the operational scoring procedure. In practice, the item calibration and the loading estimation are usually conducted in field tests before the formal operational test administration. Regarding the calibration and loading estimation procedure in HO-IRT, a number of studies can be reviewed as detailed references (Sheng & Wikle, 2007; de la Torre & Song, 2009; de la Torre & Hong, 2010; de la Torre, Song, & Hong, 2011; Huang, Wang, Chen, & Su, 2013).

Table 4

Descriptive Summary of Different Orders of Simulated Ability Parameters

Correlation	Parameter	Mean	Var	Min	Max
	General_Theta	-0.017	1.032	-3.610	3.245
Low	Subtheta_1	0.031	1.008	-2.904	3.203
	Subtheta_2	-0.013	1.030	-3.516	2.987
	Subtheta_3	-0.027	1.016	-3.096	2.897
Mixed	Subtheta_1	0.011	1.073	-3.981	3.701
	Subtheta_2	-0.009	1.049	-3.424	3.894
	Subtheta_3	0.021	1.058	-3.527	3.319
High	Subtheta_1	-0.021	0.999	-3.980	3.309
	Subtheta_2	-0.013	1.051	-3.557	3.388
	Subtheta_3	-0.020	1.041	-3.577	3.141

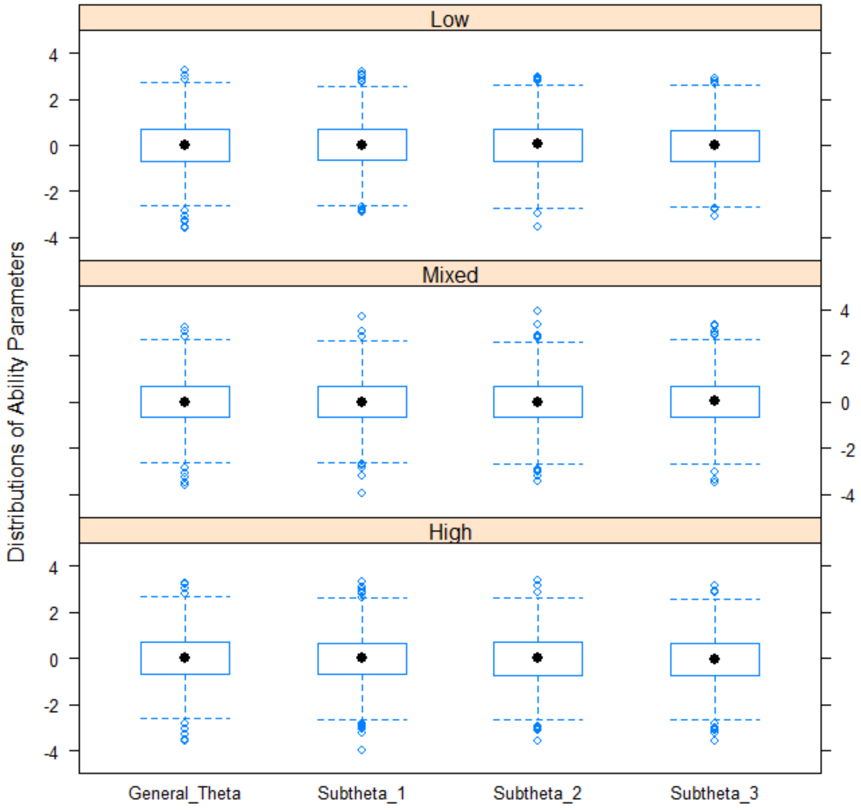

Figure 2. Boxplots of Ability Parameters Simulated in the Study.

After generating different orders of ability parameters, a UIRT model must be specified for the simulation of responses in HO-IRT. As summarized in the last section of Chapter 2, the 3PL UIRT model (Lord, 1980) was always employed in the previous studies and was also widely used in the item calibration of operational tests, which is denoted as

$$P_{ij_{(d)}}(\theta_{i(d)}) = c_{j(d)} + (1 - c_{j(d)})\frac{1}{1 + \exp[-1.7a_{j(d)}(\theta_{i(d)} - b_{j(d)})]}, \tag{65}$$

where $a_{j(d)}$ is the discrimination parameter of item *j* in subtest *d*; $b_{j(d)}$ is the difficulty parameter of item *j* in subtest *d*; and $c_{j(d)}$ is the pseudo-guessing parameter of item *j* in subtest *d*. In the study, these item parameters were directly pulled out of three operational item pools, each representing a subtest in a test battery and elaborately illustrated in the next section of this chapter. Because the 3PL UIRT model was used to calibrate these item pools in practice, it was adopted to simulate responses in the study.

Also, distinct from the other simulation studies in CAT, the responses were not to be generated during the implementation of a simulated CAT test, but an item response pool including the responses for all the subpools was established for all examinees in advance. That is, assuming that each examinee needed to answer all the items in each of the three subpools, the responses to all the items were simulated by accordingly substituting his/her subscale parameters and all the item parameters in three subpools to the 3PL UIRT model (Equation 65). Consequently, the simulated item response pool would be a $I \times (J \times D)$ matrix. The purpose of building up the item response pool in advance was to eliminate the chance of producing contradictory responses for an examinee once the same item was selected by three different algorithms (IND-UCAT, SEQ-CAT, and PC-MCAT). In this way, it was more feasible to compare subscore estimates and the usability of items in each subpool across different item selection algorithms. Note that the simulated response data is also expected to be suitable for the MIRT subscore estimation because the 3PL MIRT model is reduced to the 3PL UIRT model due to the simple structure of all items. Therefore, the model fit of MCAT to the response data is not inferior to the other methods.

In summary, this study implemented a 3 (item selection algorithms) × 5 (subscoring algorithms) × 2 (subtest length) × 3 (correlation structures of subscales) fully crossed simulation design with 90 conditions. Replications were not considered in the current study based on the fact that no replications were conducted in the literature if the overall performance of CAT algorithms across ability levels was of interest, of which the current study is such a case (Huang, Chen, & Wang, 2012; Deng, Ansley, & Chang, 2010; Barrada, Olea, Ponsoda, & Abad, 2008). The response data simulation, the item selection procedure, the subscoring procedure, the total score estimation, and the final summary analyses were all conducted by the programming language R (R Development Core Team, 2008).

Pool-Constrained MCAT (PC-MCAT) and CAT Components

The conventional MCAT is a well-recognized scoring method in the literature, which has been included in a number of comparison studies. However, under the subscoring mechanism, the item context effect may arise during the implementation of a MCAT test. It can lead examinees to a more anxious and confused testing mode when they confront the alternate item contents in a short time period (Segall, 1996). To avoid this effect, the study modified the traditional MCAT and conducted the pool-constrained MCAT (PC-MACT), of which the item selection and scoring procedures are equivalent to the traditional MCAT. The only difference was the item pool used for the item selection procedure.

To be more specific, each item in a traditional MCAT is selected from the entire item pool, which is a mixture of items from all subpools. It may often be the case that an item on math may be followed by an item on reading. The shift of item contents may make examinees more anxious and even get lost. PC-MCAT constrains the item selection to be conducted within each subpool. That is, the items are first selected from the first subpool until the fixed sublength is reached. Then the item selection moves forwards to the second subpool. It continues until the item selection procedure is completed in the last subpool. In other words, the entry of subpools for item selection in PC-MCAT is identical to the one in IND-UCAT in the study, in which each entire subpool is sequentially utilized. Aside from the entry of subpools, PC-MCAT follows the same procedures of item selection and scoring in MCAT. In terms of the shadow test, for the PC-MCAT with MAP scores and the item selection criterion of the Bayesian version of D-optimality, when the *k*th item is selected for subtest *d*, the objective function of the *k*th shadow test in Equation (23) is substituted by

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[\det(\boldsymbol{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{k-1}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}^{k-1}, u_{n_{(d)}}) + \boldsymbol{\Phi}^{-1}) \right] x_{n_{(d)}} . \qquad (66)$$

Regarding PC-MCAT, Kroehne, Goldhammer, and Partchev (2014) conducted a small-scale comparison study in which PC-MCAT (called constrained MAT (CMAT) in their study) was included for comparisons to IND-UCAT and MCAT with different content balancing techniques. In their study, the item selection algorithms of IND-UCAT and MCAT were also considered by being paired with two scoring methods of IND-UIRT and MIRT. Different from the current study, the maximum information criterion was employed as the item selection algorithm for IND-UCAT in their study. The results of their study showed favorable supports of the conventional MCAT, except for under the conditions that the items selected by IND-UCAT were scored by MIRT and the subtests in PC-MCAT were administered in some particular sequences. In the meantime, they also pointed out that future studies were required on the different configurations of subpools with different patterns of correlations among subtests. The current study is one realization of their suggestions.

To summarize, a CAT procedure is typically comprised of five key components, which are (1) the item pool, (2) the first-item entry rule, (3) the item selection criterion, (4) the scoring algorithm, and (5) the test stopping rule (Weiss & Kingsbury, 1984). With respect to the first component, an operational item pool was employed in the study, which includes three item subpools, each representing a subscale in a test battery. The test battery consists of three subtests— Language Art (LA), Applied Math (AM), and Math Computation (MC), each of which was individually used as Subtest1 to Subtest 3 in the study. After screening the items, three subpools respectively involve 281(Subtest 1 : LA), 154 (Subtest 2 : AM), and 320 (Subtest 3 : MC) dichotomously-scored items. Altogether, the study has 755 (=281 + 154 + 320) items in the entire item pool. Table 5 and Figure 3 below present the descriptive summary and the distributions of three item parameters in each subpool respectively.

Table 5

Descriptive Summary of Three Item Parameters in Each Subpool

Parameter	Subpool	Mean	SD	Min	Max
a	LA	1.124	0.414	0.355	2.229
	AM	1.029	0.325	0.415	2.003
	MC	1.140	0.415	0.335	2.446
b	LA	0.019	0.837	-2.766	2.020
	AM	0.172	1.337	-3.410	3.862
	MC	-0.447	1.355	-3.935	3.516
c	LA	0.214	0.031	0.160	0.363
	AM	0.206	0.030	0.123	0.308
	MC	0.156	0.026	0.095	0.266

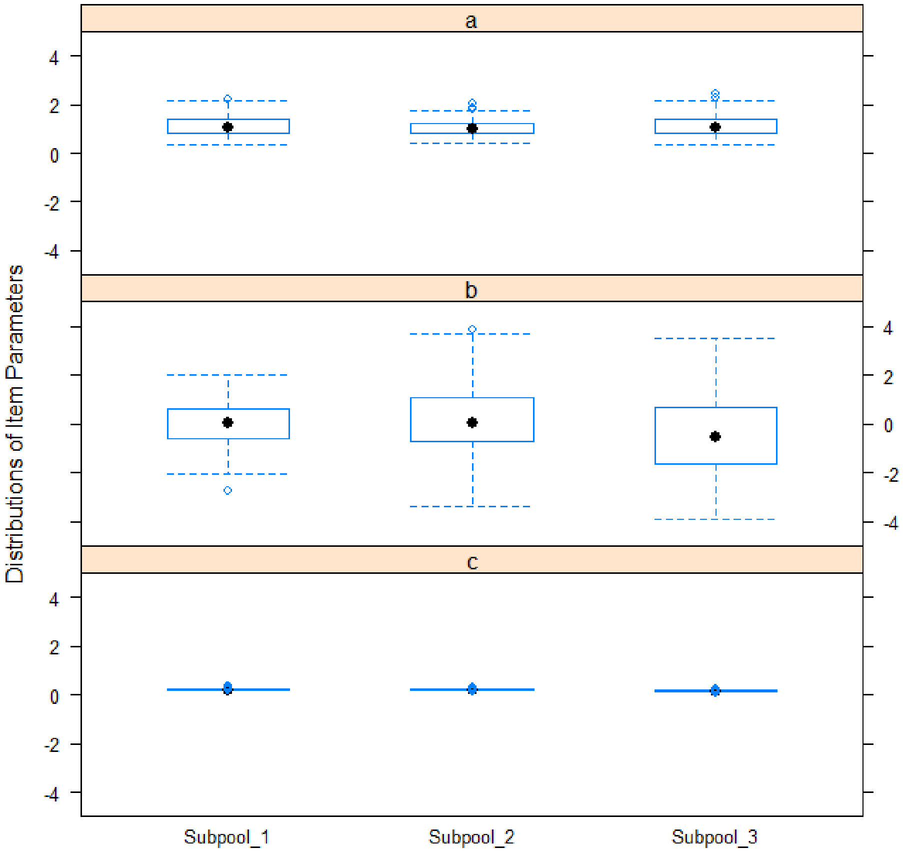

Figure 3. Boxplots of Three Item Parameters in Each Subpool.

Regarding the first-item entry rule, for PC-MCAT, the initial ability estimate for each examinee started from the average level of the first subtest, which was a scalar of 0 for LA in the study. The prior distribution for all subscales in PC-MCAT was specified as a multivariate normal distribution with a mean vector of $\left[0\ 0\ 0\right]$ and a $D \times D$ covariance matrix that is the same as Equation (64) for the corresponding simulated test conditions. Given the information provided above, the item that satisfied Equation (61) in the LA subpool was selected as the first item for all examinees in PC-MCAT. For IND-UCAT and SEQ-CAT, if the maximum information criterion is chosen as the item selection criterion, the item that provides the maximum information on the average ability level, a scalar of 0, is typically selected as the first item for each examinee. However, in the study, the maximum posterior-weighted information (MPI) criterion was adopted for item selection. For this criterion, the prior distribution for either each subtest or the entire test must be specified in advance. Once it is specified, the item that provides the maximum information integrated over the prior distribution is selected as the first item. In the study, the prior distributions of all subtests were identically assumed to be the marginal distribution of any subscale, which was a standard normal distribution ($N(0,1)$). The specification of all the prior distributions was further used in the MAP scoring procedure after the first selected item was completed and thereby the first provisional ability estimate was obtained.

Once the first item was selected and completed by an examinee, the response would update the prior distribution for IND-UCAT and SEQ-CAT. The MPI criterion and the MAP scoring algorithm for the 3PL UIRT model was subsequently implemented, for which the detailed information can be found in Sections 1, 2, and 5 of Chapter 2. Note that when the first item was to be selected in each of the subtests in IND-UCAT, the prior distribution was always a standard normal distribution ($N(0,1)$) due to the fact that no collateral information was used. For PC-MCAT, after the first selected item was completed, the first group of provisional ability parameters for all three subscales was then estimated, which was a vector with three elements. Based on these provisional ability estimates, the Bayesian version of D-optimality and the MAP scoring algorithm for the 3PL MIRT model was thereafter conducted for selecting the following items and obtaining the next few groups of provisional ability estimates, a process which is depicted in Sections 1 and 6 of Chapter 2. Note that a shadow test was calculated every time an item or a subtest was selected in the five compared methods, which is also addressed in Section 3 of Chapter 2 and the current section.

In terms of the test stopping rule, the fixed subtest length was employed for the convenience of comparison, and therefore the total test length was correspondingly fixed. As a consequence, the content constraint was imposed so that the number of items from each subscale could be balanced and the fixed subtest length could be satisfied. In addition, item security is not always a critical concern for low-stake diagnostic assessments and therefore the item exposure control was not considered in the study (van der Linden, 2010). Once the test was completed, the subscore estimates from IND-UCAT, SEQ-CAT, and PC-MCAT were all obtained. In the meantime, all the items selected by IND-UCAT, SEQ-CAT, and PC-MCAT were also recorded individually and then scored by the other four subscoring methods including two post-hoc subscoring methods, which are AUG-CAT and reSEQ-CAT.

Total Score Estimation Procedure

To make the best of the sources underlying the assessments, the study proposed a successive scoring procedure according to the structure of the higher-order IRT model, in which the test total score of individual examinees can be calculated after the subscore estimation procedure is conducted. Through the successive scoring procedure, the subscores and the total score of an examinee can be sequentially derived from one test. The successive scoring procedure is comprised of two consecutive stages. The subscoring procedures described in Chapter 2 belong to the first stage, at which point the subscores, either augmented or unaugmented, are obtained in the measurement phase of the HO-IRT model. Based on these subscore estimates, the total score estimation procedure suggested below is conducted in the structural phase of the HO-IRT model, which is regarded as the second stage.

At the first stage (the measurement phase), the subscore estimates ($\hat{\theta}_{i(1)}, \hat{\theta}_{i(2)}, \cdots,$ and $\hat{\theta}_{i(D)}$) for individual examinees are obtained through some subscoring procedure. The estimation procedure then continues to the second stage (the structural phase). As addressed in Section 7 of Chapter 2, the general ability θ_{iG} has a linear relationship with subscales $\theta_{i(d)}$, given by Equation (62). Conditional on the general ability, the distributions of the subscale parameters are correspondingly defined as $\theta_{i(d)} \mid \theta_{iG}, \lambda_d \sim N(\lambda_d \theta_{iG}, 1-\lambda_d^2)$. This association, on the other hand, illustrates that the variability of the given general ability estimate can be accounted for by the associated subscale distributions when the subscores are used for estimating the given general ability. As described previously, given θ_{iG}, the subscales, $\theta_{i(1)}, \theta_{i(2)}, \cdots$, and $\theta_{i(D)}$, are independent of each other. By assuming that the subscores estimated by the five subscoring methods are from the distribution of $\theta_{i(d)} \mid \theta_{iG}, \lambda_d \sim N(\lambda_d \theta_{iG}, 1-\lambda_d^2)$, the likelihood function is therefore obtained as

$$L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG}) = \prod_{d=1}^{D} f(\theta_{i(d)} \mid \theta_{iG}) , \tag{67}$$

where

$$f(\theta_{i(d)} \mid \theta_{iG}) = \frac{1}{\sqrt{2\pi(1-\lambda_d^2)}} \exp(-\frac{(\hat{\theta}_{i(d)} - \lambda_d \theta_{iG})^2}{2(1-\lambda_d^2)}) . \tag{68}$$

In practice, the natural logarithm of Equation (68), called the log-likelihood, is often used for convenience of computation. That is,

$$\ln L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG}) = \sum_{d=1}^{D} \ln(f(\theta_{i(d)} \mid \theta_{iG})) . \tag{69}$$

Conceptually speaking, the estimated maximum-likelihood (ML) total score $\hat{\theta}_{iG}^{ML}$ for examinee *i* is defined as

$$\hat{\theta}_{iG}^{ML} = \underset{\hat{\theta}_{iG} \in (-\infty, +\infty)}{\arg\max} (\ln L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG})) . \tag{70}$$

That is, solve

$$\frac{\partial \ln L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG})}{\partial \theta_{iG}} = 0 \tag{71}$$

for the ML total score estimate $\hat{\theta}_{iG}^{ML}$. Finally,

$$\hat{\theta}_{iG}^{ML} = \sum_{d=1}^{D} \frac{\theta_{i(d)} \lambda_d}{1 - \lambda_d^2} \Bigg/ \sum_{d=1}^{D} \frac{\lambda_d^2}{1 - \lambda_d^2} \,. \tag{72}$$

As a matter of fact, the MAP and EAP total score estimates can also be obtained by conducting the Bayesian estimation procedure, which is to integrate the likelihood function in Equation (67) to the prior distribution of the general ability. That is,

$$\begin{aligned} f(\theta_{iG} \mid \theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)}) &= \frac{L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG}) f(\theta_{iG})}{f(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)})} \\ &\propto L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG}) f(\theta_{iG}) \end{aligned} , \tag{73}$$

where $f(\theta_{iG} \mid \theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)})$ is the posterior distribution of the given θ_{iG}. The prior distribution $f(\theta_{iG})$ was assumed as a standard normal distribution in the study. Correspondingly, the EAP total score estimates are defined as

$$\hat{\theta}_{iG}^{EAP} = \int \theta_{iG} f(\theta_{iG} \mid \theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)}) d\theta_{iG} \,. \tag{74}$$

The MAP total score estimates, which are of interest in the study and are denoted as $\hat{\theta}_{iG}$ in the following chapters for consistency, are defined as

$$\hat{\theta}_{iG}^{MAP} = \underset{\hat{\theta}_{iG} \in (-\infty, +\infty)}{\arg\max} (f(\theta_{iG} \mid \theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)})) . \quad (75)$$

Similarly, solve

$$\frac{\partial \ln L(\theta_{i(1)}, \theta_{i(2)}, \cdots, \theta_{i(D)} \mid \theta_{iG})}{\partial \theta_{iG}} + \frac{\partial \ln f(\theta_{iG})}{\partial \theta_{iG}} = 0 \quad (76)$$

for the MAP total score estimates $\hat{\theta}_{iG}^{MAP}$. Finally,

$$\hat{\theta}_{iG}^{MAP} = \sum_{d=1}^{D} \frac{\theta_{i(d)} \lambda_d}{1 - \lambda_d^2} \Bigg/ \left(\sum_{d=1}^{D} \frac{\lambda_d^2}{1 - \lambda_d^2} + 1 \right). \quad (77)$$

This successive scoring procedure is applicable to both P&P tests and CAT tests. It avoids the sophisticated MCMC algorithm by assuming the regression coefficients, which can be estimated from the field tests in advance, are known, and provides total scores and subscores at one time from the same test. Moreover, the total score estimation at the second stage is fairly computable and understandable and thus holds considerable potential for future operational use. On the other hand, the successive scoring approach does not account for the estimation errors of subscore estimates in the total score estimation. However, as long as the validity and reliability of subscores are guaranteed, the estimation errors of subscores will have little impact on the accuracy of total score estimates. In this study, the proposed successive scoring approach was applied to the five compared methods so that the total scores and subscores were all provided for comparison.

Outcome Measures

As discussed in the last section of Chapter 2, the most commonly used outcome measures include the correlation, the root mean square error (RMSE), and the bias in comparison studies in the literature. In the study, these three indices were also adopted. They were separately calculated for comparisons of the recovery of total scores estimated by subscores, subscores from each subtest, and subscores from the combined three subtests (Sub_COMB). The outcome measures of Sub_COMB were utilized to evaluate the overall performance of each subscoring method on estimating the subscores across all subtests.

The correlations for the three types of scores were respectively referred to as $cor(\theta_{i(G)}, \hat{\theta}_{i(G)})$, $cor(\theta_{i(d)}, \hat{\theta}_{i(d)})$, and $cor(\theta_{i(1,2,...,D)}, \hat{\theta}_{i(1,2,...,D)})$ .The biases are individually denoted as

$$bias = \frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)}{I}; \qquad (78)$$

$$bias = \frac{\sum_{i=1}^{I}(\hat{\theta}_{i(d)} - \theta_{i(d)})}{I}; \qquad (79)$$

$$bias = \frac{\sum_{d=1}^{D}\sum_{i=1}^{I}(\hat{\theta}_{i(d)} - \theta_{i(d)})}{D \times I}. \qquad (80)$$

Moreover, the RMSEs are expressed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)^2}{I}}\ ; \tag{81}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\theta}_{i(d)} - \theta_{i(d)})^2}{I}}\ ; \tag{82}$$

$$RMSE = \sqrt{\frac{\sum_{d=1}^{D}\sum_{i=1}^{I}(\hat{\theta}_{i(d)} - \theta_{i(d)})^2}{D \times I}}\ . \tag{83}$$

After these outcome measures were calculated for all the five CAT subscoring methods in 90 conditions, they were tabled and plotted for the convenience of comparison. Strong correlations, small RMSE values, and zero biases are expected in the final results for each condition in order to justify the measurement efficiency of these subscoring methods on score estimation. Weak correlations usually reflect the large discrepancies between the score estimates and the true scores, but may not necessarily lead to large bias values. Also, correlations are negatively proportional to the RMSE values. Under the condition that no hugely discrepant cases occur between the estimates and the true values across the ability levels, positive biases could represent an overestimation pattern while negative biases represent an underestimation pattern. Compared to the RMSE and bias values in IND-UCAT, the effects of the post-hoc augmentation techniques in AUG-CAT and reSEQ-CAT as well as the subscoring advantages of PC-MCAT and SEQ-CAT can be demonstrated by relatively lower absolute values. On the other hand, the changes of the RMSE and bias values, as the factor levels investigated in the study change, also reflect how these factors impact the score estimation among the five methods.

CHAPTER IV

RESULTS

Based on the simulation design described in Chapter 3, a simulation study was correspondingly conducted and the results are presented in this chapter. For the convenience of generalization, the results of the comparison are individually illustrated from three primary perspectives, which are the three levels of the correlation structures (low, mixed, and high) as shown in Equation (64). The outcome measures (correlation, bias, and RMSE) were calculated on each type of score estimates (subscores from separate subtests, subscores from the combined three subtests (Sub_COMB), and total scores estimated from subscores) for all the conditions within each correlation structure and then all were tabled and plotted in Tables 7 to 15 and Figures 4 to 9 for straightforward visualization. They were compared in and across each simulated condition with the purpose of evaluating the performance of the five subscoring methods and the effects of the crucial factors examined in the study. Also, the depictions of collateral information in the previous chapters reflect that the collateral information utilized in the study primarily refers to the correlations among subtests. Therefore, the scoring methods that exploit the collateral information on score estimation (SEQ-CAT, PC-MCAT, reSEQ-CAT, and AUG-CAT) are generally called the correlation-based scoring methods in the study.

As mentioned previously, the item selection and scoring procedure of IND-UCAT was regarded as the baseline method for each type of score estimates, to which the other four scoring methods along with individual item selection algorithms were compared. Thus, the original values of its outcome measures are all displayed and highlighted in Tables 7 to 15. The values shown in the tables for the other methods are actually the differences from the baseline. That is, by comparison to the values of the baseline, the positive implies a larger absolute value and the negative implies a smaller absolute value. For example, the original values of correlations and RMSEs are always positive for all the methods. In Table 8, the first value in the first row is 0.915, which was highlighted for the total scores of the baseline scoring method of IND-UCAT conducted within the IND-UCAT item selection. The second value of 0.001 implies that the correlation between true total scores and total scores estimated by the SEQ-CAT scoring on the items selected by IND-UCAT was 0.001 larger than the baseline correlation of 0.915. On the other hand, the original values of biases could be either positive or negative, which reflect how much on average the scores, estimated by each method, are positively or negatively deviated from zero. For the purpose of comparison, the bias is evaluated by examining which method produces the smallest absolute values, which implies the closest to zero. For instance, the second value of the first row in Table 10 is -0.001, which means that the bias yielded by the SEQ-CAT scoring on the items selected by IND-UCAT was 0.001 closer to zero than the baseline bias. In short, by contrast to the highlighted baseline value, a positive difference value on correlation represents relative better performance whereas relative better performance reflected on bias and RMSE is represented by a negative difference value.

The item selection methods existing among the five subscoring methods were separately applied and investigated in the study, representing the item selection algorithms of IND-UCAT, SEQ-CAT, and PC-MCAT. Their performances were evaluated by comparing the outcome measures across the conditions of the three methods. Also, the difference values of the IND-UCAT subscoring in the item selections of SEQ-CAT and PC-MCAT reveal the changes of subscore estimates resulted from the use of the collateral information in the item selection algorithms rather than in the subscoring procedure. Additionally, the similarity of the items selected by these three methods was examined as well, which is shown in Table 6. The comparisons among them may provide a clue of which method could best exploit the collateral information on item selection under varied conditions.

Table 6

Percent on the Similarity of Items Selected by IND-UCAT, SEQ-CAT, and PC-MCAT

Sublength	Low	Mixed	High
10 items	.806	.760	.684
20 items	.888	.859	.811

Besides, the comparisons among the original designs of IND-UCAT, SEQ-CAT, and PC-MCAT, namely the subscoring procedure in conjunction with their own item selection algorithm, were conducted by evaluating the values on the diagonal of the corresponding submatrices within each score type in Tables 7 to 15. For instance, regarding the total score estimates in a 10-item sublength of Table 7, the elements on the diagonal of the submatrix consisting of the first three rows and the first three columns respectively represent the correlation (=0.698) of true total scores and total scores estimated by the original IND-UCAT and the difference values (= -0.001 and -0.002) of the original SEQ-CAT and the original PC-MCAT on correlation compared to the original IND-UCAT. In addition, the reSEQ-CAT scoring and the AUG-CAT scoring are both post-hoc score estimation approaches, which are applied after a conventional CAT test, such as the original IND-UCAT or the original SEQ-CAT, is completed. The comparisons between them were achieved by evaluating the values in the corresponding submatrix, such as a submatrix consisting of the first three rows and the last two columns if the example described above is also applied in this case. Also, the improvements on score estimates that these two approaches might achieve upon the three original CAT tests (the original IND-UCAT, SEQ-CAT, and PC-MCAT) were found by comparing their difference values in each row to the diagonal element in the corresponding submatrix of the three original CAT tests in the same row.

Conditions with Low Correlation Structure

In each of the conditions with the low correlation structure, the performances of the five subscoring methods were very comparable on estimating all types of scores, especially regarding the measures of correlation and RMSE. The comparability is manifested not only by the five overlapped lines in each cell of the first row of Figures 4 to 9, but also by quite small and similar difference values in each row of Tables 7, 10, and 13. For example, the difference values in the third row of Table 7 are all -0.002 for a

10-item sublength and the values in the fourth row are either 0 or 0.001for the same sublength. These values in each row are very close to or equal to zero, which implies that the scores estimated by the five scoring methods were approximate to each other and also to the highlighted baseline score within each item selection algorithm for all score types. Likewise, the comparability was further enhanced, particularly on bias, as the subtest length increased from 10 to 20. It is demonstrated by nearly no gaps between the lines in each cell of the first row of Figure 7 compared to the counterparts in Figure 6 as well as by more consistent and smaller values in each row of a 20-item sublength compared to the counterparts of a 10-item sublength in Tables 7, 10, and 13. Relatively speaking, as opposed to the difference values in Subtest 3, the measure of bias had slightly larger margins among the difference values for all the other score types, especially in a 10-item sublength. This was confirmed primarily by the fairly larger discrepancy, from 0.003 to 0.009 at the maximum across rows of Table 10, between the AUG-CAT scoring and the other scoring methods in a 10-item sublength. By contrast, the differences among the other scoring methods within each item selection were only 0 to 0.003 at the maximum in a 10-item sublength. It implies that AUG-CAT might produce, on average, larger or slightly larger positive biases within each item selection when the sublength was short and the correlations among subtests were low.

Regarding the item selection algorithms, about 81% and 89% of the total number of items (30 and 60 items) were identically selected, but perhaps not in the same sequence, by IND-UCAT, SEQ-CAT, and PC-MCAT out of the three subpools for a 10-item sublength and a 20-item sublength. It also appeared that the use of the collateral information into the item selection made no contribution to the improvements of the score estimates yielded by the five scoring methods. In some conditions, it even demolished the performances of these scoring methods, which is indicated by many negative difference values in Table 7 and many positive difference values in Tables 10 and 13 within the rows of the item selection methods of SEQ-CAT and PC-MCAT.

In short subtests (10 items), by comparing the difference values in each column within each score type in Tables 7, 10, and 13, it shows that the demolishment from the SEQ-CAT item selection algorithm was negligible due to its values approximate to the counterparts of the IND-UCAT item selection. For example, the differences between the counterparts of the two item selections on bias only ranged from 0 to 0.006. On the other hand, these small differences indicate that the adaptive selection of subtests in SEQ-CAT did not play a role in improving the score estimates when a test battery with a low correlation structure was administered. Comparatively, PC-MCAT performed the worst on the item selection for almost all score types in a 10-item sublength, especially with respect to the bias. The differences between the counterparts of IND-UCAT and PC-MCAT on bias ranged from 0.006 to 0.019 in a 10-item sublength. The big differences can also be identified from the first row of Figure 6, on which the PC-MCAT item selection positively increased the biases in all the scoring methods and the AUG-CAT scoring was most largely impacted. Consequently, the impact might lead to higher RMSE values, which are presented as some positive difference values in the rows of the PC-MCAT item selection method in Table 13. However, no matter which item selection method was adopted, the demolishment vanished as the sublength increased from 10 to

20 items. Therefore, when the correlations among subtests were low and the sublength was short, the scores estimated by the five scoring methods on the items selected by PC-MCAT were, on average, overestimated by a larger amount across the ability scales, compared to the other two selection algorithms. Under this situation, the PC-MCAT item selection could even aggravate the overestimation produced by AUG-CAT.

On the top rows of Figures 4 to 9, it shows that the large divergences of the total score estimates from the true total scores were very noticeable. Also, as the number of the items in each subtest increased, the divergences were not obviously reduced even though the subscore estimates were globally improved (see the values for Sub_COMB in Tables 7, 10, and 13 regarding a 10-item sublength versus a 20-item sublength). It is primarily manifested that as the sublength increased from 10 to 20 items, the correlations for the subscores across all the scoring methods were on average increased by 0.029 and RMSEs were on average decreased by 0.099 whereas the corresponding values for the total scores were only 0.013 and 0.015. However, the increased sublength appeared to have some comparable influence on the bias of both scores. As the sublength increased, the positive biases across all the scoring methods were, on average, reduced by 0.018 for subscore estimates and 0.017 for total score estimates. The influence was even reinforced in the PC-MCAT item selection and the AUG-CAT scoring because their biases made almost no difference to the biases from the other scoring and item selection methods in a 20-item sublength. The results described above were summarized by comparing the top row of Figure 6 to the counterparts of Figure 7 and also examining the values in Table 10.

Besides, as mentioned previously, the original IND-UCAT, SEQ-CAT, and PC-MCAT were compared by evaluating the diagonal elements in their corresponding submatrices in Tables 7, 10, and 13. It shows that when the correlation structure was low, the original IND-UCAT and SEQ-CAT exhibited very comparable performances on score estimation. The values of these two original methods in the tables demonstrate quite similar patterns regarding the three outcome measures, of which the differences were no more than 0.001 on correlation and RMSE and at most 0.004 on bias. Also, as observed above, the original PC-MCAT, on average, overestimated the subscores and total scores more than the original IND-UCAT and SEQ-CAT. Its difference values on bias to the original IND-UCAT ranged from 0.008 to 0.016. However, the distinction was eliminated to 0.005 at the maximum as the sublength increased.

Also, in terms of the measures of correlation and RMSE, reSEQ-CAT and AUG-CAT, both as the post-hoc subscore estimation method, performed in a quite similar manner on the recovery of total scores and subscores within each item selection algorithm. The differences between them were no more than 0.002 on both measures. However, unlike AUG-CAT on bias, reSEQ-CAT produced slightly lower positive biases than all the other scoring methods for all score types within each item selection, especially in a 10-item sublength. It also appeared that the combination of either method with the original IND-UCAT was much more advantageous on score estimation than their combination with the original PC-MCAT because the three measures from the former combination exhibited the best pattern among all the combinations. Additionally, as the extension of SEQ-CAT, reSEQ-CAT did not make an evident improvement on score estimation within each item selection compared to SEQ-CAT. The differences between SEQ-CAT and reSEQ-CAT were only 0.003 at the maximum on the three outcome measures.

Conditions with Mixed Correlation Structure

The mixed correlation structure in the study represents the mixture of the moderate and high correlations existing among subtests, which included the moderate correlations (0.48 and 0.40) of Subtest 1 (LA) with Subtest 2 (AM) and with Subtest 3 (MC) and also the high correlation (0.76) of Subtest 2 (AM) with Subtest 3 (MC). This type of the correlation structure occurs more often in operational tests, in which the correlations among subtests may not be at the same level. Under the conditions with the mixture of moderate and high correlations among subtests, some distinctions on the performance of score estimation were gradually presented not only among the scoring methods but also among the item selection algorithms.

Regarding the measures of correlation and RMSE within each item selection algorithm, the correlation-based scoring methods performed slightly better or better than the IND-UCAT scoring on all the score types. It is demonstrated in the middle row of Figures 4, 5, 8, and 9 and within each row of Tables 8 and 14. As the sublength increased from 10 to 20 items, their discrepancies to the IND-UCAT scoring became smaller, especially on the correlation, of which the changes, from 0 to 0.005 in a 20-item sublength of Table 8, were too small to be counted on. On RMSE, their discrepancies to IND-UCAT were reduced, but still noticeable, of which the maximum ranged from 0.004 to 0.018 in a 20-item sublength. Aside from the IND-UCAT scoring, the SEQ-CAT scoring had a poorer performance than PC-MCAT, reSEQ-CAT, and AUG-CAT in all the item selections except in its own item selection. It was especially validated on RMSE, of which its maximum difference to the other three scoring methods reached 0.026 across the item selections of IND-UCAT and PC-MCAT in a 10-item sublength and 0.015 in a 20-item sublength. However, this distinction was almost dissolved as the test proceeded, which is demonstrated by the approximate difference values between the SEQ-CAT scoring and the other three scoring methods in Subtest 3 of Tables 8 and 14. On the other hand, within its own item selection, the SEQ-CAT scoring performed very comparably to the other three scoring methods. Also, as the extension of SEQ-CAT and a post-hoc score estimation approach, the reSEQ-CAT scoring totally compensated for the weaknesses of SEQ-CAT on all score types and performed as well as PC-MCAT and AUG-CAT regardless of the item selection algorithms.

With respect to the measure of bias within each item selection algorithm, as in a 10-item sublength of the low correlation structure, the AUG-CAT scoring, on average, produced the largest positive bias for almost all score types compared to the other scoring methods, which is shown in the middle row of Figure 6. It is also shown in Table 11 that its maximum differences to the other scoring methods ranged from 0.005 to 0.011 in a 10-item sublength. From the counterparts of Figure 7, the increase of the sublength could most largely reduce the difference between AUG-CAT and the other scoring methods on bias, which is also evidently presented by comparing a 10-item sublength to a 20-item sublength in Table 11. Other than AUG-CAT, for a 10-item sublength the PC-MCAT scoring, on average, produced slightly larger positive bias than SEQ-CAT and reSEQ-

CAT and even than IND-UCAT in some conditions. Similarly as in the low correlation structure, the reSEQ-CAT scoring, on average, had the lowest positive bias, but very approximate to SEQ-CAT. When the sublength increased to 20 items, all these discrepancies among the scoring methods became negligible.

Also, apart from the IND-UCAT scoring that does not exploit the collateral information, the disparity of the correlation levels between subtests differentiated the score estimation among the subtests for the other four scoring methods. As described above, Subtest 1 had moderate correlations with the other two subtests whereas the correlation between Subtest 2 and Subtest 3 was strong. It implies that when the subscores in these three subtests were to be estimated, the amount of information from the other subtests was limited to Subtest 1, but not to Subtest 2 and Subtest 3. Therefore, compared to the baseline scoring of IND-UCAT, the improvements on score estimates by all the other scoring methods in Subtest 1 were expected to be smaller than the improvements in the other two subtests. The hypothesis is fully verified by comparing the difference values in Subtest 1 within each item selection to the counterparts of Subtest 2 and Subtest 3 in Tables 8, 11, and 14.

It is also worth noting that the difference values between Subtest 1 and the other two subtests in the SEQ-CAT item selection were not so deviated as they were in the item selections of IND-UCAT and PC-MCAT, especially regarding the measure of correlation and RMSE. This might be attributable to the adaptive sequence of subtests administered in the SEQ-CAT item selection, which was Subtest 3, Subtest 1, and then Subtest 2 for all the simulated examinees in the conditions with the mixed correlation structure. The different sequence of subtests administered in the SEQ-CAT item selection, in some degree, counterbalanced the impacts of the unbalanced correlations among subtests on score estimation and significantly contributed to the overall performance of the SEQ-CAT item selection in this case (see the difference values among the three item selections in the “Sub_COMB” score type in Tables 8, 11, and 14). Likewise, as the sublength increased, the influence of the disparity of correlations between subtests was much reduced and the differences on the improvements between Subtest 1 and the other two subtests became smaller for all the item selections.

Besides, the percent on the similarity of items selected by IND-UCAT, SEQ-CAT, and PC-MCAT was 76% for a 10-item sublength and 85% for a 20-item sublength. The evaluation of the scoring methods among the three item selection algorithms for each score type indicated that the use of the collateral information in the item selection algorithms tended to play a role in improving the score estimates, which was mostly reflected by the SEQ-CAT item selection in a 10-item sublength. It also appeared that the PC-MCAT item selection in a 10-item sublength made a big improvement on score estimation for the subtests that had high correlations to the other subtests. It is manifested by comparing the row of Subtest 1 to the rows of Subtest 2 and Subtest 3 in the PC-MCAT item selection of Tables 8, 11, and 14. On the other hand, it indicates that the moderate correlation among subtests was still not strong enough for the PC-MCAT item selection to improve the score estimation in a short subtest. As a consequence, the moderate correlations in the correlation structure neutralized the overall performance of the PC-MCAT item selection in a test battery with short subtests. It is reflected by comparing the difference values of the PC-MCAT item selection to the difference values of the IND-UCAT and SEQ-CAT item selection within the score type of Sub_COMB in Tables 8, 11, and 14.

By comparison, the overall performance of the PC-MCAT item selection was very approximate to the IND-UCAT item selection. In the score type of “Sub_COMB”, the distinction on the difference values between these two selection methods was on average 0.016, whereas the distinction between the SEQ-CAT and IND-UCAT item selections was on average 0.090 across all the three outcome measures. Comparatively speaking, the SEQ-CAT item selection was more sensitive to the moderate correlation among subtests and exhibited the best overall performance when the subtest was short in the study. As such, when the subtest was long enough, the differences among the three item selections were diminished to some extent, but were still conspicuous, particularly on bias and RMSE.

Compared to the low correlation structure, when the correlations among subtests were moderate or above, the total score estimates were very largely improved, especially regarding the measures of correlation and RMSE, although there was no such big improvements achieved for the subscore estimates from which the total scores were estimated. This is verified by comparing the middle row to the top row in Figures 4, 5, 8, and 9. Also, the increase of the number of items in each subtest not only improved the total score estimates, but also distinctly curtailed the differences on total score estimation both among the scoring methods and among the item selection algorithms, particularly in terms of the bias and RMSE within each item selection. When the subtest was short, the differentiated performances among the subscoring methods and among the item selection algorithms on the subscore estimation accordingly resulted in their differentiated performances on the total score estimation, which demonstrated a similar pattern as they were on the subscore estimation.

Specifically speaking, by evaluating the difference values in the score type of “Total” for a 10-item sublength in Tables 8, 11, and 14 and also the plots in the middle rows of Figures 4, 6, and 8, it indicates that the SEQ-CAT item selection performed the best on the total score estimation among the three selection algorithms for all the scoring methods. Within each item selection, the scoring methods of AUG-CAT and PC-MCAT, on average, produced larger positive bias of total score estimates whereas the reSEQ-CAT scoring produced the smallest. However, regarding the measures of correlation and RMSE within the item selections of IND-UCAT and PC-MCAT, the scoring methods of PC-MCAT, reSEQ-CAT, and AUG-CAT performed very comparably better than IND-UCAT and SEQ-CAT on estimating total scores.

In terms of the comparisons of the original IND-UCAT, SEQ-CAT, and PC-MCAT, the original SEQ-CAT made the best performances on almost all score types when the correlation structure consisted of two moderate correlations and one strong correlation and the subtest was short. As the subtest was spun enough, the performances of the original SEQ-CAT and the original PC-MCAT were comparably better although the latter produced relatively larger positive bias. Both of the results can be found by comparing the diagonal elements of the corresponding submatrices in Tables 8, 11, and 14. Also, like the performances in the low correlation structure, two post-hoc estimation methods, AUG-CAT and reSEQ-CAT, still performed very distinctly on bias within each item selection. However, the distinction was much reduced for a test battery with longer subtests. Otherwise, concerning the correlation and RMSE, they were very comparable on score estimation regardless of the number of items in each subtest. Additionally, for a test battery with short subtests and a mixed correlation structure, it occurred that AUG-CAT and reSEQ-CAT performed the best in conjunction with the original SEQ-CAT.

Conditions with High Correlation Structure

Strong correlations among subtests imply that more of the information collateral to the other subtests could be provided by these subtests and be utilized for the estimation of the target subscores. It allows more possibility of improvements on the score estimation. However, different scoring and item selection methods may exhibit different capabilities of making use of the information. Therefore, in the conditions with a high correlation structure, the correlation-based scoring and item selection methods became more functional and performed superiorly over the baseline method of IND-UCAT. Their performances also turned out to be more distinguishable from each other, especially on the measure of bias.

Within each item selection algorithm, all the correlation-based scoring methods performed consistently better than the IND-UCAT scoring for all the score types. Although the AUG-CAT scoring still, on average, produced the largest positive bias of all the other scoring methods in each item selection, the difference to the IND-UCAT scoring, from 0.001 to 0.007 in a 10-item sublength and from 0 to 0.001 in a 20-item sublength, became inconsequential. Also, when the correlations among subtests were all strong in a test, great changes on average bias occurred to the PC-MCAT scoring, of which the biases in the item selections of IND-UCAT and PC-MCAT were only positively larger than the reSEQ-CAT scoring that always yielded the smallest average bias among all the scoring methods. The average biases provided by the SEQ-CAT scoring were relatively moderate in the IND-UCAT and PC-MCAT item selections. They were, however, considerably reduced as the test continued.

Take the SEQ-CAT scoring in the IND-UCAT item selection for a 10-item sublength as an example. On bias, the difference of SEQ-CAT to reSEQ-CAT in Subtest 1 was 0.018 whereas the difference decreased to 0 in Subtest 3. On the other hand, associated with its own item selection, the SEQ-CAT scoring produced average bias as low as the reSEQ-CAT scoring for almost all score types. The differences to the reSEQ-CAT scoring ranged from 0 to 0.005 in a 10-item sublength whereas the differences vanished in a 20-item sublength. Also, as the sublength increased, the big distinctions on bias became small among all the scoring methods, which was indicated by the maximum difference of 0.051 for a 10-item sublength versus 0.016 for a 20-item sublength. All the results described above are accordingly presented in the bottom row of Figures 6 and 7 and in Table 12.

Despite the large discrepancies among all the scoring methods on average bias, the performances of the correlation-based scoring methods were fairly homogenous within each item selection regarding the measures of correlation and RMSE. Generally speaking, they performed uniformly better than the IND-UCAT scoring and comparably to each other, which is evidently reflected in the bottom row of Figures 4, 5, 8, and 9 and in

Tables 9 and 15. As demonstrated on bias, the performance of the SEQ-CAT scoring was comparatively weaker in the IND-UCAT and PC-MCAT item selections and was, however, remarkably improved as the test proceeded. Within its own item selection, the SEQ-CAT scoring performed better than in the other item selections, but still slightly worse than the other correlation-based scoring methods, especially regarding the total score estimates.

When the number of items in each subtest was adequately large, the differences among all the scoring methods on correlation became insignificantly small, of which the maximum values ranged from 0.001 to 0.014 for a 20-item sublength. Then in terms of RMSE, the reSEQ-CAT and PC-MCAT scoring methods performed relatively better than the other scoring methods, especially in short subtests, in which the absolute difference value of reSEQ-CAT could be as large as 0.121. Although the gaps between the IND-UCAT scoring and the other scoring methods on RMSE were shrunk for all score types when the sublength was increased from 10 to 20 items, the differences were still conspicuous and should not be ignored for each item selection method, of which the maximum values ranged from 0.007 to 0.05.

Regarding the item selection algorithms, only 68% of the items were identically, but non-synchronously, selected by IND-UCAT, SEQ-CAT, and PC-MCAT for a 10-item sublength and 81% for a 20-item sublength. The low percentage on the similarity of the selected items in short subtests may imply the large divergences among the performances of the three item selection methods on score estimation. When the subtest was short, the biases produced by all the scoring methods within the IND-UCAT item selection tended to be more deviated from each other, but on average lower than the biases within the PC-MCAT item selection. This can be verified by comparing the difference values between the IND-UCAT item selection and the PC-MCAT item selection in the score types of “Total” and “Sub_COMB” of Table 12.

The bias in the SEQ-CAT item selection was relatively more compact among all the scoring methods and was also, on average, the lowest for all the scoring methods compared to the counterparts in the item selections of IND-UCAT and PC-MCAT. This can be verified by comparing the difference values of the SEQ-CAT item selection to the ones of the IND-UCAT and PC-MCAT item selections for all score types in Table 12. However, for a test battery with long subtests, the differences among the scoring methods and among the item selections on bias were simultaneously reduced for all score types. These results could be found by comparing the bottom row of Figure 6 to the counterpart of Figure 7 as well as by evaluating the difference values in a 10-item sublength versus in a 20-item sublength in Table 12.

As for the measure of correlation, the differences among the three item selection methods ranged from 0.001 to 0.009 for all the scoring methods, which are shown for the score types of “Total” and “Sub_COMB” in Table 9. It indicates that there were, on average, almost no big differences among the three item selection methods across the two sublengths. On the other hand, with respect to RMSE, the large difference values occurred to the item selection methods of SEQ-CAT and PC-MCAT, especially for the correlation-based scoring methods in short subtests. This can be detected by comparing the difference values in the item selections of SEQ-CAT and PC-MCAT to the ones in the IND-UCAT item selection for a 10-item sublength of all score types in Table 15. As the sublength increased, the differences among the three item selection methods decreased, of which the maximum absolute difference value across the item selection methods of SEQ-CAT and PC-MCAT decreased from 0.121 in a 10-item sublength to 0.059 in a 20-item sublength compared to the IND-UCAT item selection. Also, the maximum difference between the SEQ-CAT and PC-MCAT item selections occurred in Subtest 1 and decreased from 0.094 in a 10-item sublength to 0.054 in a 20-item sublength. All of the values above are derived from Table 15.

Besides, by comparing the three outcome measures on the "Total" score type to the counterparts on the "Sub_COMB" score type in Tables 9, 12, and 15, it is of great interest to find that all the scoring methods demonstrated a slightly better performance on the total score estimation than they did on the subscore estimation when the correlations among subtests were all strong. Generally speaking, for a test battery with short subtests, the PC-MCAT and reSEQ-CAT scoring methods performed relatively better than the other scoring methods in each item selection algorithm regarding the measures of correlation and RMSE. As the sublength increased, the differences among the scoring methods, from 0.001 to 0.008, became negligible.

On the other hand, the increase of sublength could not totally eliminate, but reduced the large disparities among the scoring methods on bias, of which the maximum dropped from 0.017 in a 10-item sublength to 0.008 in a 20-item sublength. Among all the scoring methods, the reSEQ-CAT scoring always produced the lowest average biases of total score estimates, from 0 to 0.017, across the three item selections and two sublengths whereas the AUG-CAT scoring produced the largest biases, from 0.004 to 0.032, comparably to the IND-UCAT scoring as they did in the subscore estimation. Comparatively, all the scoring methods in the PC-MCAT item selection performed consistently better than they did in the other two item selections with regard to the measures of correlation and RMSE of total score estimates. However, regarding the measure of bias, they performed the best in the SEQ-CAT item selection and the differences among them were also more condensed. As the sublength increased, all the discrepancies among the item selections on the three measures, from 0.001 to 0.005 in a 20-item sublength, became insignificantly small.

As for the three original scoring methods, the original SEQ-CAT and the original PC-MCAT both performed better than the original IND-UCAT on the score estimation, which is revealed by comparing the diagonal elements in the corresponding submatrices in Tables 9, 12, and 15. It also suggests that the original PC-MCAT should be employed for estimating all types of scores in either a 10-item sublength or a 20-item sublength. The reason for the use of the original PC-MCAT is because the original SEQ-CAT exhibited a weaker performance on the total score estimation, particularly in a test battery with short subtests. The weaknesses were primarily manifested by the measures of correlation and RMSE of total score estimates, for example, 0.009 lower on correlation and 0.032 higher on RMSE in a 10-item sublength compared to the original PC-MCAT.

In addition, of the two post-hoc score estimation methods, the performance of the reSEQ-CAT scoring exhaustively exceeded the AUG-CAT scoring for all score types in all item selections of both sublengths. It was further validated when both methods were implemented jointly with the original PC-MCAT. Relatively speaking, the AUG-CAT scoring performed better in conjunction with the original SEQ-CAT than with the original PC-MCAT. However, the improvements still could not surpass the improvements achieved by the reSEQ-CAT scoring within the item selections of SEQ-CAT and PC-MCAT. It is also very interesting to find that the reSEQ-CAT scoring combined to the original IND-UCAT sometimes performed comparably to or even better than some correlation-based scoring methods in the other two item selections, such as the AUG-CAT and SEQ-CAT scoring methods in the PC-MCAT item selection in a 10-item sublength. As always, the score estimates from the SEQ-CAT scoring were improved to a large degree by the reSEQ-CAT scoring, particularly when the sublength was short.

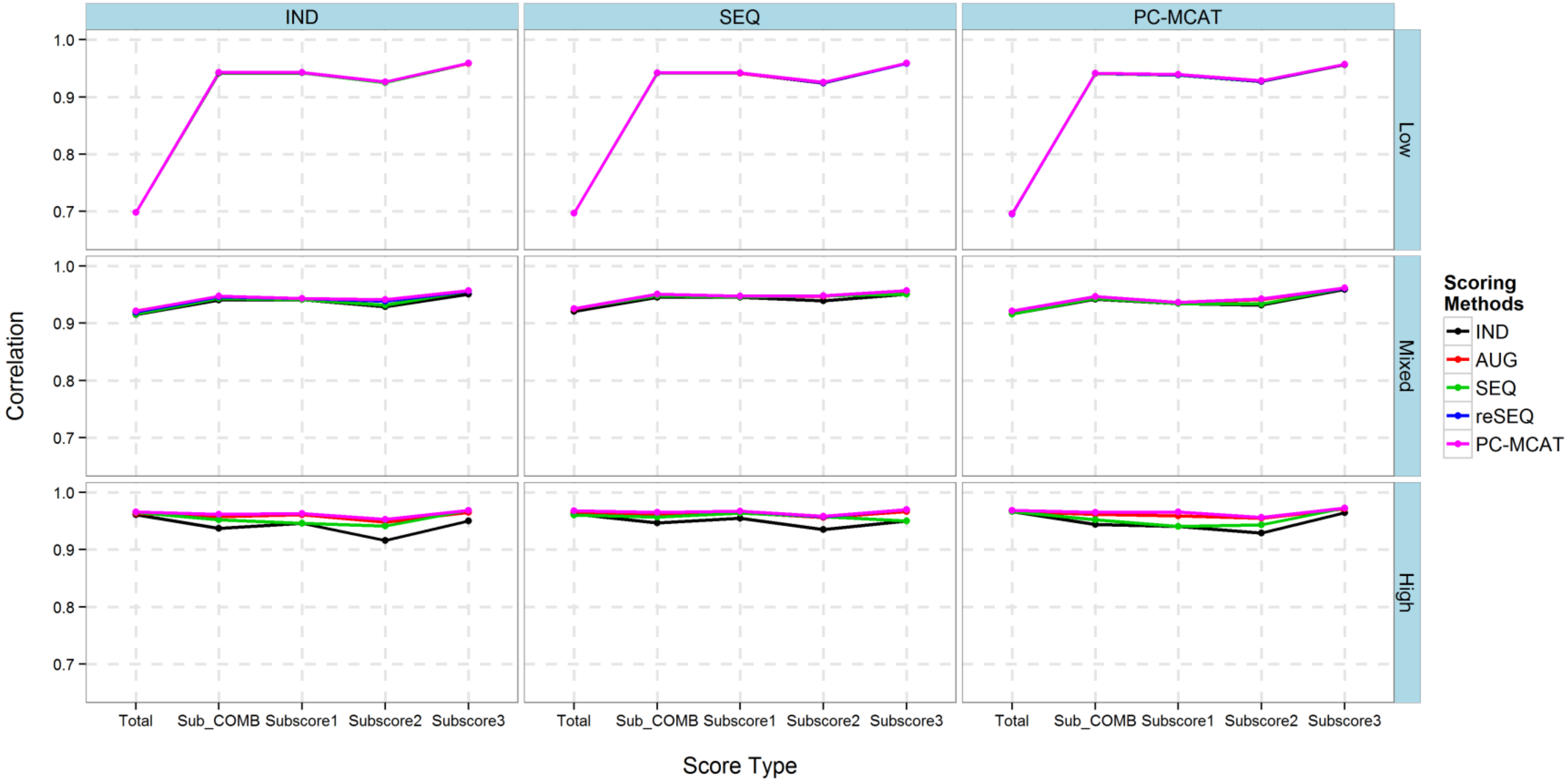

Figure 4. Correlation between θ and $\hat{\theta}$ for All the Conditions with A 10-Item Sub-length.
Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

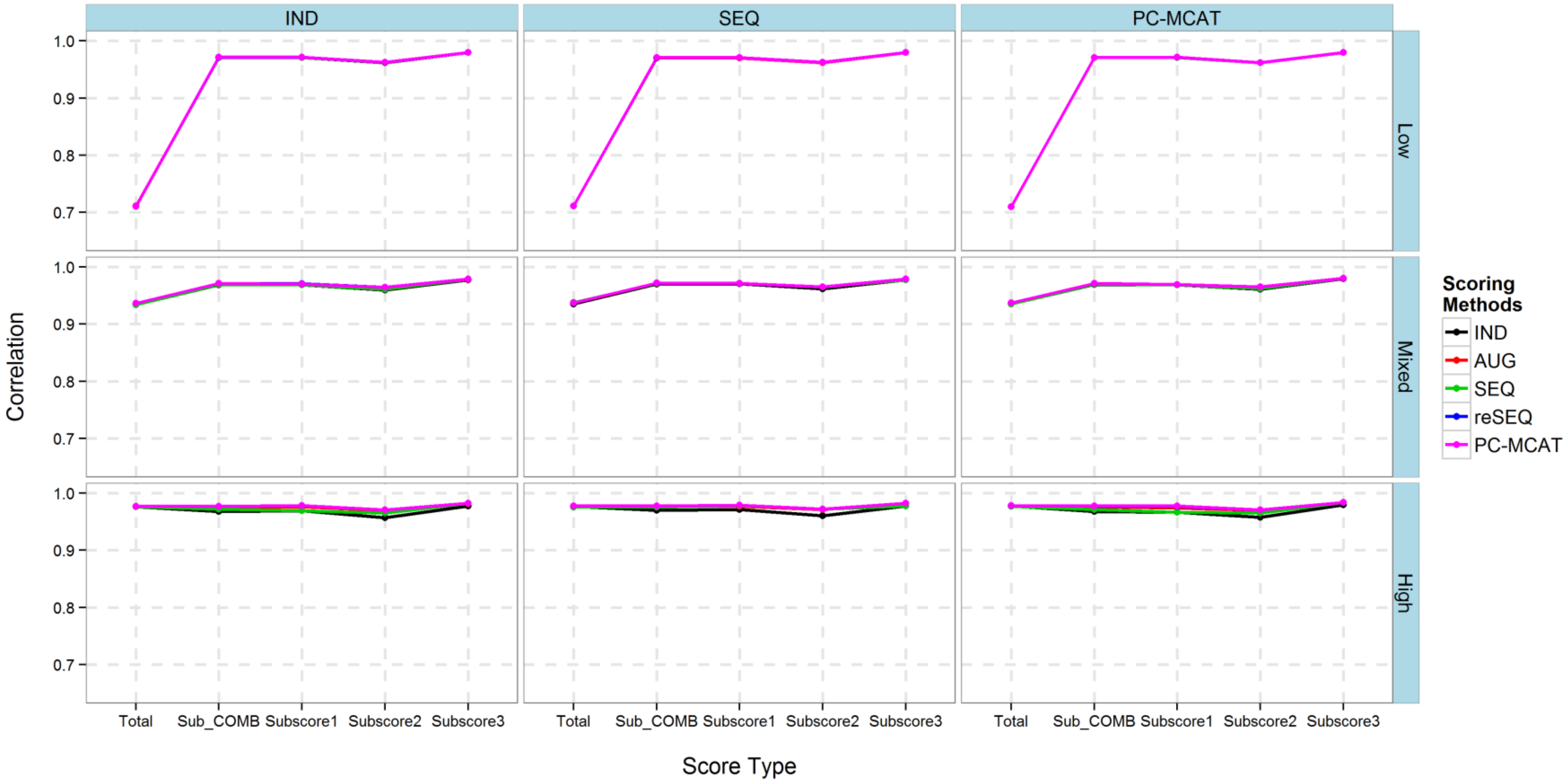

Figure 5. Correlation between θ and $\hat{\theta}$ for All the Conditions with A 20-Item Sub-length.

Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

Table 7

Correlation (Difference Values) between θ and $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.698	.000	.000	.000	.000	.711	.000	.000	.000	.000
	SEQ-CAT	-.001	-.001	-.001	-.001	-.001	.000	.000	.000	.000	.000
	PC-MCAT	-.002	-.002	-.002	-.002	-.002	-.001	-.001	-.001	-.001	-.001
Sub_COMB	IND-UCAT	.942	.000	.001	.001	.001	.971	.000	.000	.000	.000
	SEQ-CAT	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	PC-MCAT	-.001	-.001	.000	.000	-.001	.000	.000	.000	.000	.000
Subtest 1	IND-UCAT	.942	.000	.001	.001	.000	.971	.000	.000	.000	.000
	SEQ-CAT	-.001	.000	.000	.000	.000	-.001	.000	.000	.000	-.001
	PC-MCAT	-.004	-.004	-.003	-.003	-.003	.000	.000	.000	.000	.000
Subtest 2	IND-UCAT	.925	.000	.002	.001	.002	.962	.001	.001	.000	.000
	SEQ-CAT	-.001	.000	.001	.000	.001	.000	.000	.000	.000	.000
	PC-MCAT	.002	.003	.003	.003	.003	.000	.000	.000	.000	.000
Subtest 3	IND-UCAT	.958	.001	.001	.001	.000	.979	.000	.000	.000	.000
	SEQ-CAT	.000	.000	.001	.001	.000	.000	.000	.000	.000	.000
	PC-MCAT	-.002	-.001	-.001	-.001	-.001	.000	.000	.000	.000	.000

Note. The highlighted values are the original values of the correlation between θ and $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on correlation; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 17, they were caused by rounding errors.

Table 8

Correlation (Difference Values) between θ and $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.915	.001	.006	.004	.006	.933	.000	.003	.003	.003
	SEQ-CAT	.005	.010	.010	.010	.010	.002	.004	.004	.004	.004
	PC-MCAT	.000	.001	.006	.006	.005	.001	.002	.004	.004	.004
Sub_COMB	IND-UCAT	.940	.003	.007	.006	.006	.969	.001	.003	.003	.002
	SEQ-CAT	.005	.008	.010	.010	.010	.001	.002	.003	.003	.003
	PC-MCAT	.001	.002	.006	.006	.006	.001	.001	.003	.003	.003
Subtest 1	IND-UCAT	.941	.000	.002	.002	.002	.969	.000	.001	.001	.001
	SEQ-CAT	.004	.005	.006	.006	.005	.002	.002	.002	.003	.002
	PC-MCAT	-.007	-.007	-.005	-.005	-.005	-.001	-.001	.000	.000	.000
Subtest 2	IND-UCAT	.929	.003	.012	.010	.011	.960	.001	.005	.005	.005
	SEQ-CAT	.010	.019	.019	.019	.019	.002	.005	.005	.005	.005
	PC-MCAT	.003	.005	.014	.014	.013	.001	.002	.005	.005	.005
Subtest 3	IND-UCAT	.951	.005	.006	.005	.005	.977	.002	.002	.002	.001
	SEQ-CAT	.000	.000	.006	.006	.006	.000	.000	.002	.002	.001
	PC-MCAT	.008	.010	.011	.010	.011	.002	.003	.003	.003	.003

Note. The highlighted values are the original values of the correlation between θ and $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on correlation; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 18, they were caused by rounding errors.

Table 9

Correlation (Difference Values) between θ and $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.961	.004	.005	.005	.002	.976	.000	.001	.001	.000
	SEQ-CAT	.002	-.001	.007	.007	.004	.000	.000	.001	.001	.001
	PC-MCAT	.006	.005	.008	.008	.006	.001	.001	.002	.002	.001
Sub_COMB	IND-UCAT	.937	.015	.025	.024	.020	.967	.004	.009	.009	.008
	SEQ-CAT	.009	.020	.028	.028	.025	.002	.008	.010	.010	.009
	PC-MCAT	.007	.015	.028	.028	.024	.000	.004	.010	.010	.008
Subtest 1	IND-UCAT	.946	.000	.017	.017	.015	.969	.000	.009	.009	.007
	SEQ-CAT	.009	.018	.021	.021	.019	.002	.009	.010	.010	.008
	PC-MCAT	-.006	-.006	.020	.020	.013	-.003	-.003	.008	.008	.005
Subtest 2	IND-UCAT	.916	.025	.037	.037	.031	.957	.008	.014	.014	.013
	SEQ-CAT	.019	.041	.042	.042	.040	.004	.015	.015	.015	.015
	PC-MCAT	.013	.027	.040	.040	.039	.001	.008	.014	.014	.013
Subtest 3	IND-UCAT	.950	.018	.018	.018	.015	.977	.005	.005	.005	.004
	SEQ-CAT	.000	.000	.020	.020	.016	.000	.000	.005	.005	.005
	PC-MCAT	.014	.022	.022	.022	.021	.002	.006	.006	.006	.006

Note. The highlighted values are the original values of the correlation between θ and $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on correlation; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 19, they were caused by rounding errors.

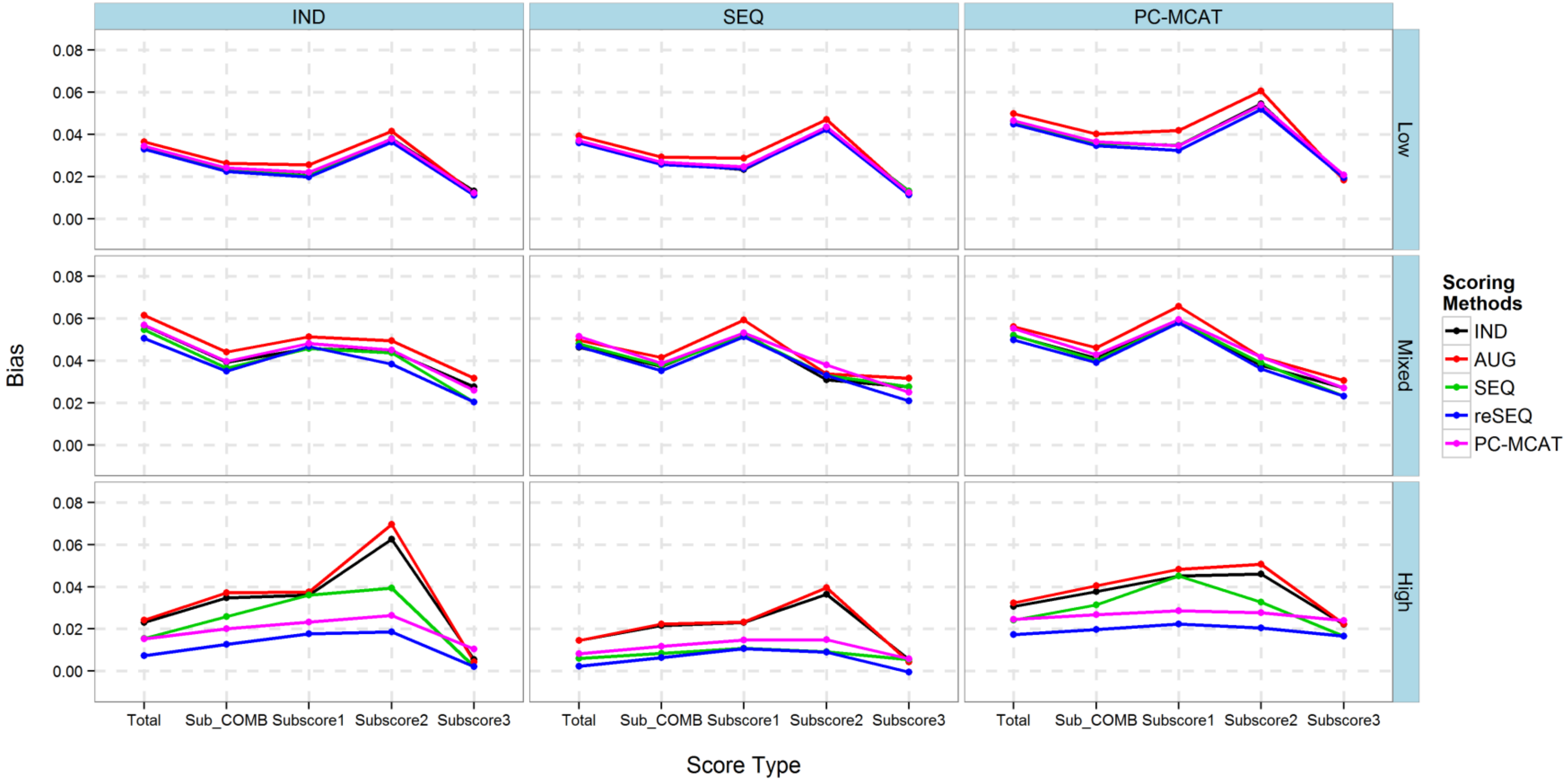

Figure 6. Bias of $\hat{\theta}$ for All the Conditions with A 10-Item Sub-length.

Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the bias; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

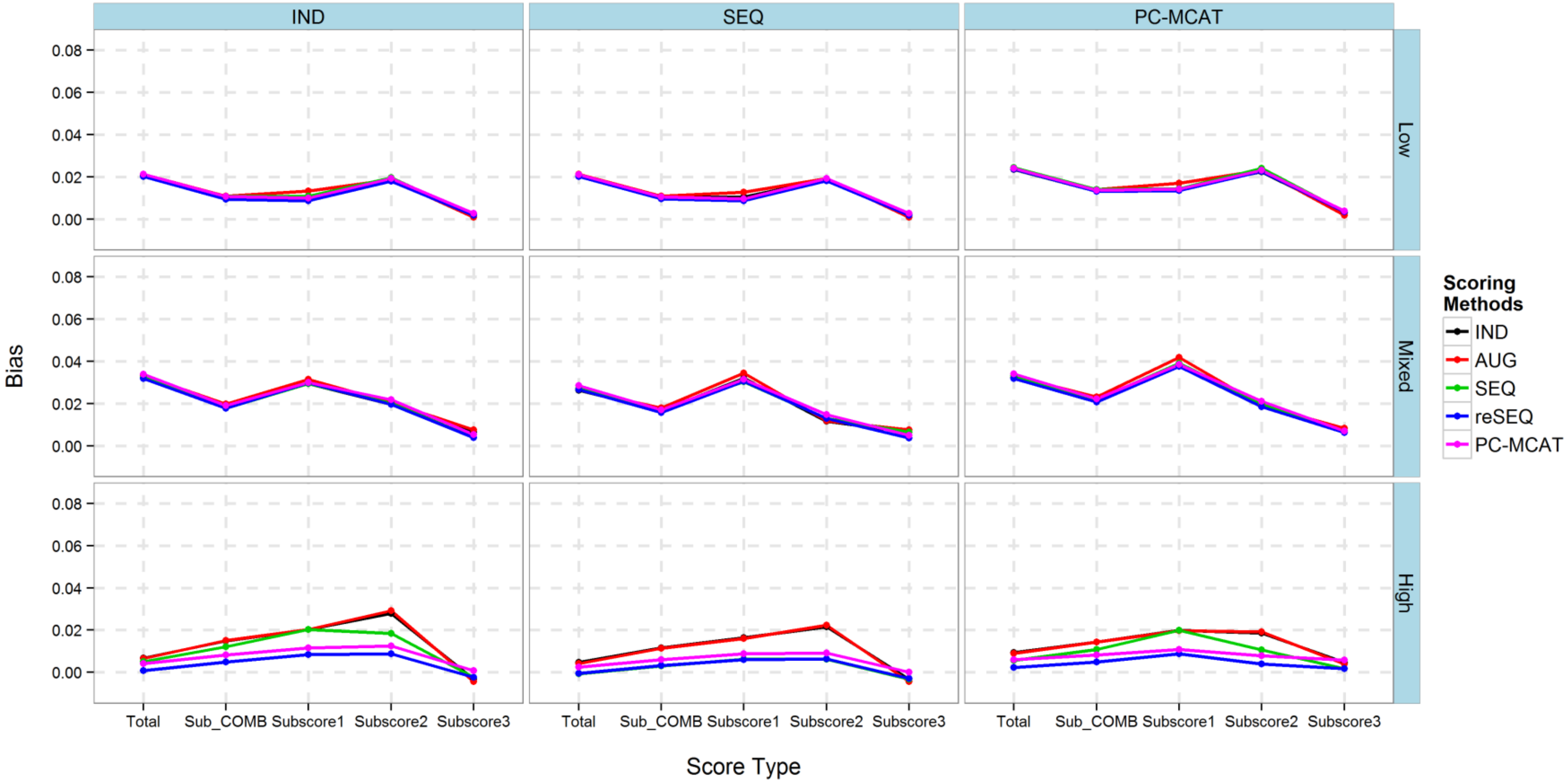

Figure 7. Bias of $\hat{\theta}$ for All the Conditions with A 20-Item Sub-length.

Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the bias; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

Table 10

Bias (Difference Values) of $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.034	-.001	.000	-.001	.002	.021	.000	.000	-.001	.000
	SEQ-CAT	.002	.002	.003	.002	.005	.000	.000	.000	-.001	.000
	PC-MCAT	.012	.012	.012	.010	.015	.003	.003	.003	.003	.003
Sub_COMB	IND-UCAT	.024	-.001	.000	-.001	.003	.010	.000	.000	-.001	.001
	SEQ-CAT	.003	.003	.003	.002	.005	.000	-.001	.000	-.001	.001
	PC-MCAT	.012	.012	.013	.011	.016	.003	.004	.003	.003	.004
Subtest 1	IND-UCAT	.020	.000	.002	-.001	.005	.011	.000	-.001	-.002	.002
	SEQ-CAT	.003	.003	.004	.003	.008	.000	-.001	-.001	-.002	.002
	PC-MCAT	.014	.014	.014	.012	.021	.003	.003	.003	.003	.006
Subtest 2	IND-UCAT	.038	.000	.000	-.002	.004	.018	.001	.001	-.001	.000
	SEQ-CAT	.005	.004	.006	.004	.009	.000	.000	.001	.000	.001
	PC-MCAT	.017	.016	.016	.014	.023	.004	.006	.005	.004	.005
Subtest 3	IND-UCAT	.013	-.002	-.001	-.002	-.001	.002	.000	.001	.000	-.001
	SEQ-CAT	.000	.000	-.001	-.002	-.001	.000	.000	.001	.000	-.001
	PC-MCAT	.006	.006	.008	.006	.005	.001	.002	.002	.002	.000

Note. The highlighted values are the original values of the bias of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on bias; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 20, they were caused by rounding errors.

Table 11

Bias (Difference Values) of $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.057	-.002	.000	-.006	.005	.032	.000	.001	-.001	.001
	SEQ-CAT	-.010	-.009	-.005	-.010	-.007	-.006	-.005	-.004	-.006	-.006
	PC-MCAT	-.005	-.004	-.001	-.007	-.001	.000	.001	.002	-.001	.001
Sub_COMB	IND-UCAT	.039	-.003	.000	-.004	.005	.018	-.001	.001	-.001	.001
	SEQ-CAT	-.002	-.002	.000	-.004	.002	-.002	-.002	-.001	-.003	-.001
	PC-MCAT	.002	.001	.004	.000	.007	.003	.003	.004	.002	.005
Subtest 1	IND-UCAT	.046	.000	.002	.001	.006	.029	.000	.001	.001	.002
	SEQ-CAT	.007	.007	.008	.006	.014	.003	.001	.002	.001	.005
	PC-MCAT	.013	.013	.014	.012	.020	.010	.010	.009	.008	.013
Subtest 2	IND-UCAT	.044	-.001	.001	-.006	.005	.020	.000	.002	.000	.001
	SEQ-CAT	-.014	-.011	-.006	-.011	-.011	-.008	-.007	-.005	-.007	-.008
	PC-MCAT	-.007	-.006	-.003	-.008	-.003	-.001	.000	.001	-.001	.000
Subtest 3	IND-UCAT	.027	-.007	-.002	-.007	.004	.006	-.003	-.001	-.003	.001
	SEQ-CAT	.000	.000	-.003	-.007	.004	.000	.000	-.001	-.003	.001
	PC-MCAT	.000	-.004	.000	-.004	.003	.001	.000	.001	.000	.002

Note. The highlighted values are the original values of the bias of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on bias; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 21, they were caused by rounding errors.

Table 12

Bias (Difference Values) of $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.023	-.008	-.008	-.016	.001	.007	-.002	-.003	-.006	.000
	SEQ-CAT	-.008	-.017	-.015	-.021	-.008	-.002	-.006	-.004	-.006	-.003
	PC-MCAT	.008	.001	.002	-.006	.009	.003	-.001	-.001	-.005	.002
Sub_COMB	IND-UCAT	.035	-.009	-.015	-.022	.002	.015	-.003	-.007	-.010	.000
	SEQ-CAT	-.013	-.026	-.023	-.028	-.012	-.003	-.012	-.009	-.012	-.004
	PC-MCAT	.003	-.003	-.008	-.015	.006	-.001	-.004	-.007	-.010	-.001
Subtest 1	IND-UCAT	.036	.000	-.013	-.018	.002	.020	.000	-.009	-.012	.000
	SEQ-CAT	-.013	-.025	-.021	-.025	-.013	-.004	-.014	-.012	-.014	-.004
	PC-MCAT	.009	.009	-.007	-.014	.012	-.001	-.001	-.010	-.012	-.001
Subtest 2	IND-UCAT	.063	-.023	-.036	-.044	.007	.028	-.009	-.015	-.019	.001
	SEQ-CAT	-.026	-.053	-.048	-.054	-.023	-.006	-.022	-.019	-.022	-.006
	PC-MCAT	-.016	-.030	-.035	-.042	-.012	-.009	-.017	-.020	-.024	-.009
Subtest 3	IND-UCAT	.005	-.003	.005	-.003	-.001	-.003	-.001	-.002	-.001	.001
	SEQ-CAT	.000	.000	.000	-.005	-.001	.000	.000	-.003	.000	.001
	PC-MCAT	.017	.011	.018	.011	.017	.001	-.002	.002	-.002	.001

Note. The highlighted values are the original values of the bias of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on bias; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation. ; If some discrepancies occurred between this table and Table 22, they were caused by rounding errors.

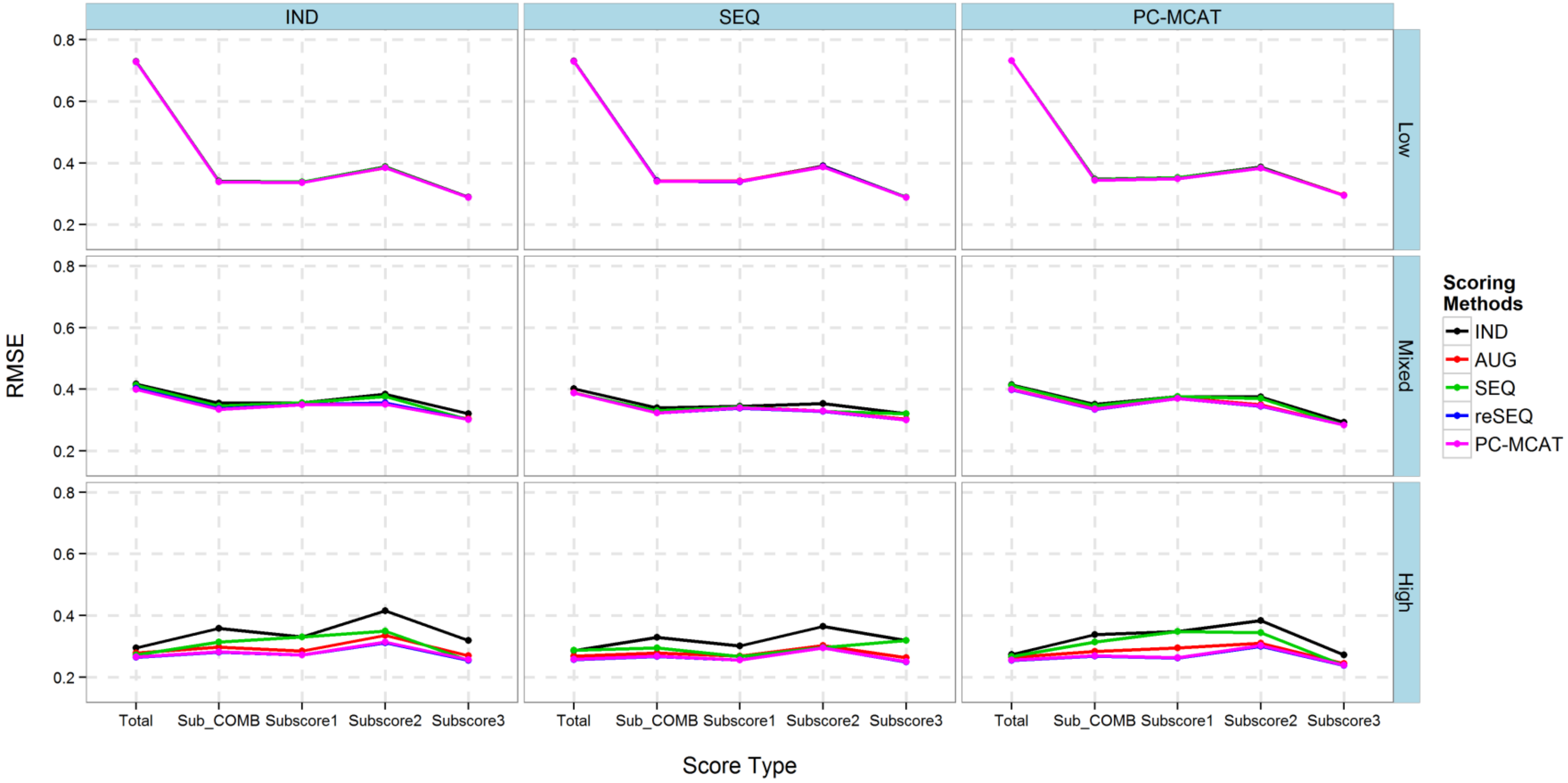

Figure 8. RMSE of $\hat{\theta}$ for All the Conditions with A 10-Item Sub-length.
Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of RMSE; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

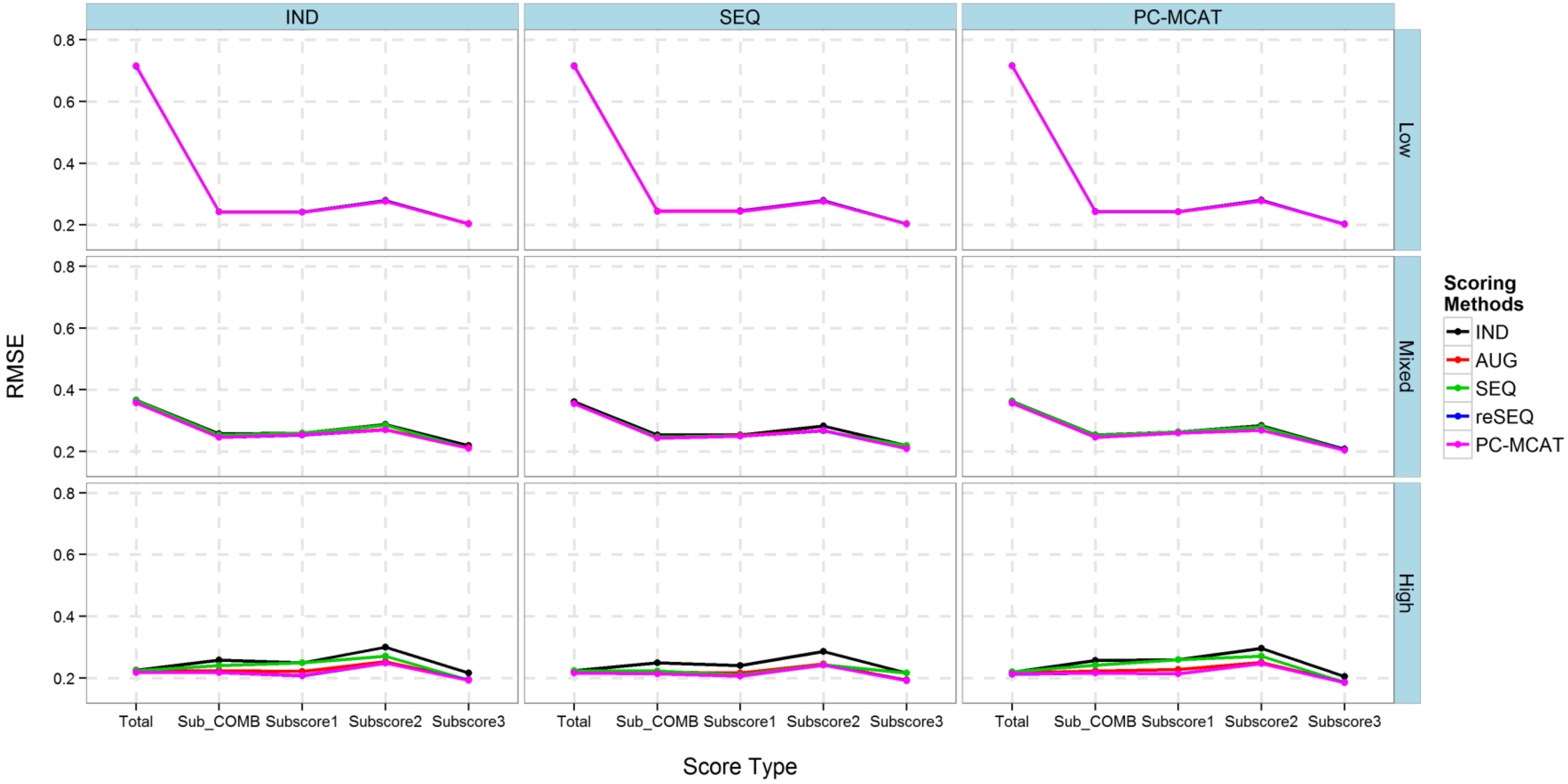

Figure 9. RMSE of $\hat{\theta}$ for All the Conditions with A 20-Item Sub-length.

Note. The three columns represent the three item selection algorithms; The three rows represent the three correlation structures; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of RMSE; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

Table 13

RMSE (Difference Values) of $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.729	-.001	-.001	-.001	-.001	.715	.000	.000	.000	.000
	SEQ-CAT	.001	.000	.000	.000	.000	.000	-.001	.000	.000	-.001
	PC-MCAT	.003	.002	.002	.002	.002	.001	.001	.001	.001	.001
Sub_COMB	IND-UCAT	.341	-.001	-.003	-.003	-.002	.244	-.001	-.002	-.002	-.001
	SEQ-CAT	.001	.000	-.001	-.001	.000	.001	.000	-.001	.000	.000
	PC-MCAT	.006	.005	.003	.003	.004	.000	.000	-.001	-.001	-.001
Subtest 1	IND-UCAT	.339	.000	-.002	-.002	-.001	.243	.000	-.002	-.002	-.001
	SEQ-CAT	.002	.001	.001	.000	.002	.003	.002	.001	.002	.003
	PC-MCAT	.014	.014	.011	.011	.013	.000	.000	-.001	-.001	.000
Subtest 2	IND-UCAT	.388	-.001	-.004	-.004	-.004	.279	-.002	-.002	-.002	-.001
	SEQ-CAT	.002	-.001	-.001	-.001	.000	.000	-.002	-.002	-.002	-.001
	PC-MCAT	-.002	-.004	-.005	-.005	-.004	.001	.000	-.001	.000	.000
Subtest 3	IND-UCAT	.290	-.002	-.002	-.002	-.001	.204	-.001	-.001	-.001	.000
	SEQ-CAT	.000	.000	-.002	-.002	-.001	.000	.000	-.001	-.001	.000
	PC-MCAT	.006	.005	.004	.005	.006	-.002	-.002	-.002	-.002	-.002

Note. The highlighted values are the original values of the RMSE of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on RMSE; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 23, they were caused by rounding errors.

Table 14

RMSE (Difference Values) of $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.417	-.003	-.017	-.014	-.015	.366	-.001	-.009	-.009	-.008
	SEQ-CAT	-.015	-.028	-.029	-.030	-.027	-.005	-.012	-.011	-.012	-.012
	PC-MCAT	-.003	-.005	-.018	-.019	-.016	-.004	-.004	-.010	-.010	-.010
Sub_COMB	IND-UCAT	.354	-.008	-.019	-.018	-.017	.257	-.003	-.011	-.011	-.009
	SEQ-CAT	-.014	-.024	-.031	-.031	-.029	-.004	-.010	-.013	-.014	-.012
	PC-MCAT	-.004	-.008	-.019	-.019	-.017	-.003	-.005	-.011	-.011	-.010
Subtest 1	IND-UCAT	.356	.000	-.006	-.007	-.005	.260	.000	-.005	-.006	-.003
	SEQ-CAT	-.011	-.014	-.016	-.017	-.014	-.006	-.009	-.010	-.010	-.008
	PC-MCAT	.021	.021	.015	.015	.016	.004	.004	.000	.000	.002
Subtest 2	IND-UCAT	.384	-.007	-.033	-.027	-.029	.288	-.003	-.018	-.017	-.016
	SEQ-CAT	-.030	-.055	-.054	-.055	-.054	-.006	-.020	-.020	-.020	-.020
	PC-MCAT	-.008	-.013	-.038	-.039	-.034	-.005	-.008	-.019	-.020	-.019
Subtest 3	IND-UCAT	.320	-.019	-.018	-.019	-.016	.219	-.008	-.009	-.008	-.006
	SEQ-CAT	.000	.000	-.021	-.020	-.017	.000	.000	-.009	-.010	-.006
	PC-MCAT	-.028	-.036	-.036	-.036	-.036	-.011	-.014	-.015	-.014	-.014

Note. The highlighted values are the original values of the RMSE of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on RMSE; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 24, they were caused by rounding errors.

Table 15

RMSE (Difference Values) of $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	.295	-.023	-.029	-.031	-.017	.226	-.002	-.008	-.008	-.005
	SEQ-CAT	-.010	-.008	-.038	-.039	-.027	-.002	-.002	-.010	-.010	-.007
	PC-MCAT	-.022	-.028	-.040	-.041	-.031	-.006	-.005	-.012	-.013	-.010
Sub_COMB	IND-UCAT	.358	-.043	-.076	-.077	-.060	.258	-.018	-.040	-.040	-.033
	SEQ-CAT	-.028	-.063	-.090	-.090	-.079	-.009	-.035	-.044	-.044	-.039
	PC-MCAT	-.020	-.043	-.088	-.090	-.074	-.002	-.016	-.041	-.041	-.035
Subtest 1	IND-UCAT	.331	.000	-.058	-.059	-.046	.250	.000	-.041	-.042	-.028
	SEQ-CAT	-.030	-.063	-.075	-.076	-.062	-.010	-.040	-.044	-.044	-.034
	PC-MCAT	.018	.018	-.067	-.069	-.036	.010	.010	-.036	-.036	-.022
Subtest 2	IND-UCAT	.416	-.066	-.102	-.104	-.080	.300	-.030	-.053	-.053	-.047
	SEQ-CAT	-.051	-.119	-.120	-.121	-.114	-.014	-.058	-.059	-.059	-.055
	PC-MCAT	-.032	-.072	-.113	-.115	-.106	-.004	-.030	-.053	-.054	-.050
Subtest 3	IND-UCAT	.319	-.064	-.062	-.064	-.049	.217	-.023	-.024	-.023	-.022
	SEQ-CAT	.000	.000	-.069	-.070	-.055	.000	.000	-.025	-.025	-.024
	PC-MCAT	-.046	-.081	-.079	-.081	-.075	-.012	-.031	-.031	-.031	-.030

Note. The highlighted values are the original values of the RMSE of $\hat{\theta}$ estimated by the baseline method of the IND-UCAT scoring in the IND-UCAT item selection algorithm; The other values are the differences between the scoring method and the baseline method on RMSE; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 25, they were caused by rounding errors.

CHAPTER V

CONCLUSIONS AND DISCUSSION

In this study, a simulation study was conducted with the primary purposes of comparing five subscoring methods on score estimation under variable simulated conditions in computerized adaptive tests. Among the five subscoring methods, the IND-UCAT scoring ignores the correlations among subtests, whereas the other four methods (SEQ-CAT, PC-MCAT, reSEQ-CAT, and AUG-CAT) are the correlation-based scoring methods, which implies that they capitalize on the correlation information in the scoring procedure. By manipulating the sublengths, the correlation structures, and the item selection algorithms, more comparable, pragmatic and systematic testing scenarios were created for comparison purposes, so that the comprehensive conclusions could be reached through comparisons in the study. Also, some particular features presented in the study may benefit or impede the exertions of some subscoring methods on score estimation. Therefore, these features should be pointed out in this chapter as attentive references for future studies and practical applications.

Conclusions Regarding the Research Questions

Based on the purposes of the study, five research questions were correspondingly raised in Chapter 1 and were exhaustively resolved in Chapter 4. Through comparisons, the results of the study demonstrated unambiguous answers to all the research questions and also provided evident support to the performances of the correlation-based scoring methods over the IND-UCAT scoring when the correlation structure was moderate or above. In addition, given different correlation structures, these correlation-based scoring methods exhibited their own weaknesses and strengths in estimating scores. Therefore, their applications will largely depend on their feasibility and efficiency to the demands and objectives of test users.

First, for the low correlation structure, by comparison to the baseline method of IND-UCAT, the utilization of the information on the correlations among subtests does not provide a visible improvement on subscore estimates for either the correlation-based subscoring methods or the correlation-based item selection methods. Conversely, it may, on average, lead to larger positive bias of subscore estimates, unless the number of items in each subtest is adequately large. This finding is especially validated for AUG-CAT, the PC-MCAT item selection, and their combination. As a method of exploiting the collateral information, the SEQ-CAT scoring and item selection perform very approximately to the baseline methods of the IND-UCAT scoring and item selection no matter which sublength (10 or 20 items) is applied. Under this situation, considering the ease of implementation in practice, the original IND-UCAT should be considered for use when the subscores are to be estimated. Also, in order to achieve an acceptable accuracy of subscore estimates for all the subscoring methods, at least 20 items in each subtest are needed. Regarding the total score estimation, as the second stage of the successive scoring procedure proposed in the study, the approach of estimating total scores is not suggested when the correlations among subtests are low. The aberrant values on all the three outcome measures mirror the very large discrepancy of the total score estimates from the true total scores regardless of the magnitude of the sublengths.

Second, when the correlation structure is comprised of two moderate correlations and one high correlation, the correlation-based item selection and scoring methods, to some extent, exhibit their advantages of estimating scores over the baseline method of IND-UCAT. The subscores and total scores estimated by these methods are all improved. The improvements are particularly pronounced for these scoring methods conducted within the SEQ-CAT item selection. On the other hand, the AUG-CAT and PC-MCAT scoring, the PC-MCAT item selection, and their combinations still, on average, produce larger positive bias for all score types. However, the discrepancies to the other methods on bias are largely reduced or even eliminated as the sublength increases. Also, when the correlations among subtests are moderate or above, the total score estimates are remarkably improved to a relatively acceptable level of accuracy, especially for a test battery with sufficient items in each subtest. Given the results, the original SEQ-CAT or the combination of the SEQ-CAT item selection and the PC-MCAT scoring are recommended not only for subscore estimation but also for total score estimation. If time and cost allow the post-hoc estimation procedure, the original SEQ-CAT and the reSEQ-CAT scoring could be jointly conducted for the best score estimates for a test with the mixed correlation structure. Again, longer subtests are always preferred in these conditions.

Last but not least, for the conditions with a high correlation structure, the increased amount of the collateral information among subtests leads to the large discrepancy between the IND-UCAT scoring and the other scoring methods on the score estimation. In the meantime, the disparities among the correlation-based scoring methods also become more distinct, especially regarding the measure of bias and RMSE. Generally speaking, the scoring methods of reSEQ-CAT and PC-MCAT perform better than the other scoring methods in all the three item selections. The performance of the SEQ-CAT scoring is not ideal on score estimation, especially on the total score estimation. Although the differences among the subscoring methods across the item selections could be reduced by the increased sublength, they are still noticeable in a 20-item sublength. The total score estimation is greatly achieved and is even globally better than the subscore estimation for both sublengths when the correlation structure is high. Another interesting finding points to some correlation-based scoring methods (e.g. PC-MCAT or reSEQ-CAT) within the IND-UCAT item selection, of which the quality of the score estimates are even better than the one obtained by some correlation-based scoring methods (e.g. AUG-CAT or SEQ-CAT) conducted within the SEQ-CAT or PC-MCAT item selection. As a matter of fact, this phenomenon may also occur in some conditions with a mixed correlation structure. Based on all the findings above, the original PC-MCAT and the combination of the PC-MCAT scoring and the SEQ-CAT item selection are suggested for both the subscore estimation and the total score estimation. If the post-hoc score estimation is allowed, the reSEQ-CAT scoring in conjunction with the original SEQ-CAT is strongly recommended. If the complexity of the implementation is an issue, the reSEQ-CAT jointly conducted with the original IND-UCAT can be considered for reasonable score estimates.

Some Thoughts on Subscore Estimation

As indicated above, the IND-UCAT subscoring in the IND-UCAT item selection (the original IND-UCAT) does not exploit the collateral information in the subscoring and item selection procedure, and therefore its performance on subscore estimation is not impacted by the levels of the correlation structures in all the conditions. Consequently, the subscores estimated by IND-UCAT should be quite similar across the three correlation structures. If some divergences occur, it is mostly attributable to the sampling and estimation errors. In the comparisons of the study, the effects of these errors have been ruled out by calculating the difference values for the other subscoring methods to the highlighted baseline values within their own correlation structure on the same sample.

Also, as one of the correlation-based subscoring methods, the SEQ-CAT scoring takes advantages of the correlation information in the subscoring procedure, which greatly facilitates its subscore estimation. As described in Section 5 of Chapter 2, all the correlation information among subtests is reflected by the joint distribution of all subscale parameters in the second-level model of SEQ-CAT. However, at the very beginning of the subtest selection and subscoring procedure, only the relevant joint marginal distribution is used and updated to select the items from the optimally selected subtests, which implies that there is less correlation information exploited in the SEQ-CAT subscoring at the early stage. It is particularly validated to the first selected subtest. In fact, when the first subtest is to be selected in SEQ-CAT, the respective marginal distribution for each subtest is actually its initial prior distribution ($f(\theta_{(d)}) = N(0,1)$), which is the same across all the subtests in the study. As a result, the subscoring procedure in the first selected subtest is equivalent to the subscoring procedure of IND-UCAT in the same subtest, which indicates that the subscores estimated by IND-UCAT and SEQ-CAT are totally identical in that subtest. It is obviously manifested by some 0 difference values between SEQ-CAT and IND-UCAT in some subtests of Tables 7 to 15.

Consequently, compared to the other correlation-based subscoring methods in the study, the SEQ-CAT subscoring does not utilize all the collateral information in at least one subtest, which partly impairs its overall performance on subscore estimation, especially when the number of subtests is small and the correlation structure is high. However, as the test proceeds, the relevant joint marginal distribution in SEQ-CAT is expanded by the later selected subtests and is also updated by the responses in the later selected subtests. The subscores in the later subtests accordingly become more and more accurate. For the first selected few selected subtests, the increase of the sublength can curtail the discrepancy between SEQ-CAT and the other correlation-based subscoring methods on the subscore estimates. On the other hand, the newly-developed reSEQ-CAT subscoring method (W. J. van der Linden, personal communication, July 30th, 2013) possesses the capacity of utilizing all the collateral information to estimate the subscores in all the subtests, and therefore exhibited the best performance among all the subscoring methods, particularly when the correlation structure was moderate or above.

Furthermore, it is not hard to find that all the subscoring methods tend on average to overestimate the subscores, which is demonstrated by almost all the positive bias values shown in Tables 19 to 21 for each condition. In order to better examine the subscore estimates and explain the phenomenon, the conditional biases were also investigated and calculated within each segment over the general ability scale and each subscale. Twelve segments were divided with almost a 0.5 unit in-between on each scale, namely, $(-\infty, -2.5]$, $(-2.5, -2]$, …, $(2.5, +\infty)$, and all the values of conditional biases were plotted in Figures 11 to 16 of APPENDIX D. Figures 11 to 16 show that the biases produced by all the five subscoring methods were deviated far from 0 for the extreme abilities on the two ends, unless the subpools were well constructed such as Subpool 3. Subpool 3 included the largest number of items and more items measuring the extreme abilities as shown in Table 5 and Figure 3. Therefore the average biases in Subtest 3 presented in Tables 10 to 12 were the lowest and most positively close to 0 among the three subtests. Due to the lack of items measuring the negative extreme abilities in Subpool 1 and Subpool 2, as shown in Figures 11 to 16, the biases produced by all the subscoring methods in these two subtests were much more deviated from 0 on the negative end than on the positive end over the ability scale, and therefore most of the biases are shown as positive values in Tables 10 to 12. When the correlation structure was low, the distinctions among all the subscoring methods were too small (at most 0.009) to be presented in Figures 11 and 12. As the correlations among subtests increased, the distinctions became more evident on the two ends, particularly for the high correlation structure, which is shown in Figures 15 and 16.

When the correlation structure was high, Figures 15 and 16 show that the biases yielded by the subscoring methods of IND-UCAT, SEQ-CAT, and AUG-CAT were relatively larger than the biases of PC-MCAT and reSEQ-CAT on both ends. However, as indicated above, the SEQ-CAT subscoring in the first selected subtest is equivalent to the IND-UCAT subscoring. Therefore, the large biases on both ends yielded by SEQ-CAT mostly occur in the first administered or selected subtest. As the test proceeds and the correlation information is involved in the subscore estimation, the biases produced by the SEQ-CAT scoring become comparably as small as the biases of PC-MCAT and reSEQ-CAT on both ends, particularly within its own item selection. Also, based on the results of this study, the increase of the sublength improves the bias and the conditional bias, but does not change the general pattern demonstrated by the five scoring methods.

Overall, among the five subscoring methods, IND-UCAT represents the implementation of multiple UCAT subtests, which are totally independent to each other and administered in a fixed and prespecified sequence. In IND-UCAT, the prior information for the MAP scoring procedure in each subtest ($\hat{\theta}_{i(d)}$) is only associated with its own subscale ability distribution ($f(\theta_{(d)})$), so no information on the correlations among subtests is involved in the scoring procedure. By contrast, the MAP scores in all the subtests of PC-MCAT are simultaneously derived from the updated prior distribution. The initial prior distribution ($f(\theta_{(1)}, \theta_{(2)}, ..., \theta_{(D)})$) is typically a multivariate normal distribution ($\text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$), which involves the information on the entire correlation structure, and is consecutively updated by the responses from each subtest. The sequence of subtests to be administered is also fixed and predetermined in PC-MCAT.

SEQ-CAT reflects a two-level empirical Bayes scoring approach, of which the relevant joint marginal distribution (i.e. $f(\theta_{(d^2)}, \theta_{(d^1)})$) regarding the current selected subtest ($\theta_{(d^2)}$) is updated by the responses in all the preceding subtests ($f(\boldsymbol{u}_{i(d^1)} | \theta_{(d^1)})$), and then the updated posterior predictive distribution ($f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)})$) is used as the prior information to estimate the MAP scores in the selected subtest. In other words, other than the responses to the items within its own subtest, the MAP scores in each subtest of SEQ-CAT are impacted not only by the correlation information among subtests but also by the responses in the previous subtests. The information provided by both regards also influences the entry of subtests into the administration, so the sequence of subtest administrations is adaptive to the performance of an examinee in the previous subtests. However, as opposed to PC-MCAT, the entire correlation structure is not exploited along with the responses in all the previous subtests for the SEQ-CAT scoring procedure until the last subtest is selected, which impedes the efficiency of SEQ-CAT to some extent, especially when the correlation structure is moderate or above.

The three methods mentioned above are all just-in-time subscoring methods, which are implemented during the administration of a CAT test. Apart from them, the other two subscoring methods (AUG-CAT and reSEQ-CAT) possess post-hoc augmentation algorithms and are implemented after a conventional CAT test. AUG-CAT is conducted on the subscores, which are estimated by the conventional CAT scoring procedure, whereas reSEQ-CAT is conducted on the items, which are selected and completed by a conventional CAT test. More precisely, the augmentation in AUG-CAT is achieved by a multivariate regression function, of which the regression coefficients regarding all the subscore estimates are determined by both the reliabilities of subscore estimates and the correlations between subtests. If the subscore estimates obtained from a CAT test are

Bayesian version, they need to be converted as unaugmented subscores by excluding the prior information from these Bayesian score estimates.

Distinct from AUG-CAT, the augmentation in reSEQ-CAT is achieved by reimplementing the MAP scoring procedure with the prior distribution reformulated by all the items that are selected by a conventional CAT test. Compared to SEQ-CAT, reSEQ-CAT is a fully Bayesian estimation approach by assuming that the prior distribution of each subscale, which is obtained by marginalizing the other subscales out of the updated joint distribution, is known. Also, for a simple-structure test battery, the final posterior density function of each subscale in reSEQ-CAT is essentially proportional to the marginal density function of the corresponding subscale given all the responses in PC-MCAT, which is demonstrated as below. For convenience, all the formulae below are expressed based on the design of the study. However, they can be easily generalized to the cases with more simple-structure subscales. By Equation (2), the final posterior distribution of PC-MCAT is written as

$$f(\theta_{(1)},\theta_{(2)},\theta_{(3)} \mid \boldsymbol{u}_{i(1)},\boldsymbol{u}_{i(2)},\boldsymbol{u}_{i(3)}) = \frac{L(\boldsymbol{u}_{i(1)} \mid \theta_{(1)})L(\boldsymbol{u}_{i(2)} \mid \theta_{(2)})L(\boldsymbol{u}_{i(3)} \mid \theta_{(3)})f(\theta_{(1)},\theta_{(2)},\theta_{(3)})}{C_A}, \quad (84)$$

where C_A represents the normalizing constant in PC-MCAT. Then the marginal distribution of Subscale 1 is denoted as

$$\begin{aligned} f(\theta_{(1)} \mid \boldsymbol{u}_{i(1)},\boldsymbol{u}_{i(2)},\boldsymbol{u}_{i(3)}) &= \iint f(\theta_{(1)},\theta_{(2)},\theta_{(3)} \mid \boldsymbol{u}_{i(1)},\boldsymbol{u}_{i(2)},\boldsymbol{u}_{i(3)})d\theta_{(1)}d\theta_{(2)} \\ &= \frac{L(\boldsymbol{u}_{i(1)} \mid \theta_{(1)})\iint L(\boldsymbol{u}_{i(2)} \mid \theta_{(2)})L(\boldsymbol{u}_{i(3)} \mid \theta_{(3)})f(\theta_{(1)},\theta_{(2)},\theta_{(3)})d\theta_{(2)}d\theta_{(3)}}{C_A}. \end{aligned} \quad (85)$$

As such, the marginal distribution of the other subscales can be obtained in the same manner.

On the other hand, reSEQ-CAT reimplements the MAP scoring procedure using the reformulated prior distribution for each subscale. For instance, the prior distribution of Subscale 1, which was selected as the second subtest for most of the examinees in SEQ-CAT, is given by Equation (49) as

$$f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)}, \boldsymbol{u}_{i(d^3)}) = \frac{\iint f(\theta_{(d^1)}, \theta_{(d^2)}, \theta_{(d^3)}) L(\boldsymbol{u}_{i(d^1)} \mid \theta_{(d^1)}) L(\boldsymbol{u}_{i(d^3)} \mid \theta_{(d^3)}) d\theta_{(d^1)} d\theta_{(d^3)}}{C_B}, \quad (86)$$

where C_B represents the normalizing constant for $f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)}, \boldsymbol{u}_{i(d^3)})$, and d^2, d^1 and d^3 correspond to Subtest 1, Subtest 3, and Subtest 2 respectively for most examinees in PC-MCAT. Then the posterior distribution of Subtest 1 in reSEQ-CAT is expressed as

$$\begin{aligned} & f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)} \boldsymbol{u}_{i(d^2)}, \boldsymbol{u}_{i(d^3)}) \\ & \quad = \frac{L(\boldsymbol{u}_{i(d^2)} \mid \theta_{(d^2)}) \iint L(\boldsymbol{u}_{i(d^1)} \mid \theta_{(d^1)}) L(\boldsymbol{u}_{i(d^3)} \mid \theta_{(d^3)}) f(\theta_{(d^1)}, \theta_{(d^2)}, \theta_{(d^3)}) d\theta_{(d^1)} d\theta_{(d^3)}}{C_C C_B}, \end{aligned} \quad (87)$$

where C_C is the normalizing constant for $f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)} \boldsymbol{u}_{i(d^2)}, \boldsymbol{u}_{i(d^3)})$. Similarly, the posterior distribution of the other subscales in reSEQ-CAT can be denoted in the same fashion. By comparing Equation (85) to Equation (87), it is obviously shown that these two functions are proportional, which implies that the maximum of these two distributions is expected to point to the same solution.

In addition, for better evaluating the performances of PC-MCAT, the comparison between PC-MCAT and MCAT was also conducted, of which the results are shown in APPENDIX E. In principle, there are no differences regarding the scoring procedure between PC-MCAT and MCAT. Both of them adopt the MAP scoring procedure in a multivariate distribution. However, the items selected by both methods might be different due to the different pool(s) used, which may therefore lead to different subscore estimates. This regard is discussed in the fourth section of this chapter.

Some Thoughts on Total Score Estimation

The total score estimation approach suggested in the study is established on the theoretical principle of the likelihood function, given that all the loadings/regression coefficients (λ_d) are known and all the subscores are given. In Chapter 4, the results showed that the total score estimates were astonishingly deviated from the true total scores when the correlation structure was low. The MAP total scores are adopted in the study and denoted as Equation (77). Algebraically, Equation (77) addresses that the MAP total score estimates are primarily determined by the magnitudes of the loadings and the subscore estimates. The loadings play a role of weights for each subscore estimate in Equation (77) for the calculation of a total score estimate.

In the low correlation structure, the subscore estimates yielded by all the subscoring methods were not sufficiently accurate as shown in Chapter 4. The values of all the loadings (from 0.45 to 0.55) were small, reflecting the weak associations between the general ability and all the subscales. As the primary sources for total score estimation, there is no doubt that the inaccurate subscore estimates in conjunction with small loadings would lead to the huge departures of total score estimates. Also, small loadings indicate large variances of subscores given the Equation of $\theta_{i(d)} \mid \theta_{iG}, \lambda_d \sim N(\lambda_d \theta_{iG}, 1-\lambda_d^2)$ and accordingly lead to much larger standard errors of total score estimates. There were only three subtests for total score estimation in the study. When the number of subtests is not sufficient, the true distribution of the general ability is hard to approximate.

Based on the same logic, for the mixed correlation structure that included two moderate correlations (0.40 and 0.48) and one strong correlation (0.76), the loadings became as large as 0.80 and 0.95, and the subscore estimates were also improved compared to the ones in the low correlation structure. Some large loadings reflect the strong associations between the general ability and some of the subscales, and imply better use of the information from subtests by giving more weights to the subscore estimates. In particular, a large weight (0.80) was given to the subscores in Subtest 3 that were the most accurately estimated among the three subtests. As a consequence, although they were still not as good as the subscore estimates regarding the three outcome measures, the total score estimates were dramatically improved to an acceptable level by comparison to the ones in the low correlation structure.

For the high correlation structure, all the loadings were larger than 0.90 and one of them (=0.98) was even close to 1. It implies that very strong correlations exist between the general ability and all the subscales and the largest amount of information from subtests can contribute to the total score estimation. Furthermore, in addition to the most improved subscore estimates, the large loadings indicate the very small variance of the subscores and also very small standard errors of total score estimates. The true distribution of the general ability can be the most approximately estimated even if the number of subtests is small. In this optimal condition, the total score estimates could be statistically and algebraically improved to the fullest, of which the three outcome measures exhibited an even better pattern than the ones for the subscore estimates in the study.

Table 16

Three Outcome Measures Regarding the Total Score Estimates When True Subscores Are Applied

	Correlation	Bias	RMSE
Low	0.722	0.011	0.702
Mixed	0.963	0.014	0.274
High	0.988	-0.001	0.154

In order to better demonstrate the total score estimation under the structure of higher-order IRT model, the true values of all subscores were applied in Equation (77), and the three outcome measures were accordingly calculated, which are shown in Table 16 above. In this way, the measurement errors produced in the measurement phase of the HO-IRT model were ruled out. By comparing these outcome measures to the corresponding values in Tables 7 to 15, the total scores estimated by true subscores were considerably improved on bias in all the three correlation structures, which implies that the majority of biases in total score estimates result from the biases produced in subscore estimates. In terms of correlation and RMSE shown in Table 16, there was a huge jump (improvement) from the low correlation structure to the mixed correlation structure. Holding subscores true in both structures, it is concluded that the aberrant total score estimates in the low correlation structure are largely attributed to the small loadings between subscales and the general ability rather than the estimation errors of subscores.

On the other hand, the combination of three subtests possesses three times the number of items in each subtest. It is commonsense in IRT that when item parameters are known, the more items used for estimating ability parameters, the more accurate these parameter estimates should be. Oddly, regardless of the sublengths in the low and mixed correlation structures of the study, the three outcome measures regarding subscore estimates always demonstrated better patterns than the values regarding total score estimates. The increase of the sublength from 10 to 20 items greatly improved the subscore estimates far more than the total score estimates. These findings also aligned with the results of de la Torre and Song's study (2009), which were found in the similar conditions. This oddity, in fact, emphasizes the properties of the HO-IRT model. First, the structural phase of the HO-IRT model reflects the causal relations among the unobservable latent traits, in which the total score estimation approach suggested in the study is conducted. Second, the subscore estimates conditional on a total score are assumed to be the observed samples from the distribution of $\theta_{i(d)} \mid \theta_{iG}, \lambda_d \sim N(\lambda_d \theta_{iG}, 1-\lambda_d^2)$. The distribution is primarily determined by the magnitudes of the loadings, as is demonstrated by the results in Table 16.

When the loadings and the number of observations are small, the distribution of the total score is poorly approximated based on the principle of likelihood function. As illustrated above, the number of observations refers to the number of subscores for each examinee in the study. Moreover, the increase of the sublength can improve the total score estimates to some extent, but not as straightforwardly and significantly as the increase of the number of subtests in a test. Therefore, to improve the total score estimates in the low or mixed correlation structure, it is suggested that more subtests, rather than more items in each subtest, should be included in a test battery, which can more approximately estimate the true distribution of the total score. This finding was also justified in de la Torre and Song's study (2009) by their conclusion that "in improving the overall ability estimates, the number of dimensions had greater impact than the number of items" (p. 627).

Additionally, as suggested previously, in the optimal condition/high correlation structure, the SEQ-CAT subscoring is not recommended on score estimation, especially on total score estimation, even within its own item selection when the number of subtests is small. That is because one of the three subscores that are used to estimate the total score in SEQ-CAT is identical to the subscore of IND-UCAT in the same subtest, which is estimated without the collateral information, and therefore is less accurate compared to the subscores estimated by the other correlation-based subscoring methods. This weakness of the SEQ-CAT subscoring becomes relatively detrimental when the correlations among subtests are all high. In the meantime, the small number of subtests allows more credits to be granted to each subscore on the total score estimation, and thus one inaccurate subscore may largely deviate the accuracy of the total score estimate.

Under the same optimal condition, it appeared that the PC-MCAT subscoring and item selection methods demonstrated stronger capabilities for taking advantage of the collateral information in the subscoring and item selection procedure, compared to their performances in the low and mixed correlation structures and SEQ-CAT. However, the impact of the weakness of SEQ-CAT is expected to be less critical as the number of subtests increases. One reason is that the large number of subtests can contribute most approximately to the true total score distribution. The other reason is that more subtests are expected to neutralize the negative impacts of the first few selected subtests in SEQ-CAT. Also, adding more items in SEQ-CAT would not be considered as one of the solutions to efficiently and significantly offset the negative impact.

Some Thoughts on Item and Subtest Selection Algorithms

As opposed to the two post-hoc estimation methods (AUG-CAT and reSEQ-CAT), the other three subscoring methods have their own item selection algorithm. IND-UCAT ignores the collateral information existing among subtests and merely uses the prior information ($f(\theta_{(d)})$) regarding its own distribution for the adaptive MPI item selection. SEQ-CAT gradually adds more and more collateral information to the prior distribution ($f(\theta_{(d^1)})$ first, then $f(\theta_{(d^2)} \mid \boldsymbol{u}_{i(d^1)})$, and then $f(\theta_{(d^3)} \mid \boldsymbol{u}_{i(d^1)}, \boldsymbol{u}_{i(d^2)})$, and so on) for the MPI item selection as the test proceeds. As for PC-MCAT, the item selection conducted in the study adopted the Bayesian version of D-optimality for each subtest, which involves the prior covariance matrix that reflects the associations among all subtests underlying $f(\theta_{(1)}, \theta_{(2)}, ..., \theta_{(D)})$.

Similarly in the SEQ-CAT subscoring, the items selected by SEQ-CAT in the first selected subtest are identical to the ones selected by IND-UCAT in the same subtest, and therefore the subscores estimated by SEQ-CAT and IND-UCAT in that subtest are totally identical. As discussed above, this weakness of SEQ-CAT could impede the overall performance of the SEQ-CAT subscoring and item selection, especially for the high correlation structure. By comparison, the PC-MCAT item selection exhibited the best performance on score estimation in the high correlation structure, which might lead to the conclusion that it possesses the best capability of utilizing the collateral information in item selection, when the correlations among subtests are sufficiently strong.

However, a weakness may not actually be "weak" in some conditions. In the low correlation structure, the performance of the SEQ-CAT item selection on subscore estimation was not largely demolished by comparison to the PC-MCAT item selection, given the results in Chapter 4 that its performance was very comparable to the IND-UCAT item selection. It believes that the subscore estimation in SEQ-CAT employs less collateral information in the first few subtests and it most likely dilutes the demolishment.

Additionally, the SEQ-CAT item selection appears more sensitive to the moderate correlations among subtests compared to the PC-MCAT item selection. In the mixed correlation structure with two moderate correlations and one strong correlation, Subtest 1 had a relatively weaker and moderate correlation with the other two subtests in the study. For Subtest 1, the PC-MCAT item selection demonstrated a similar pattern as it did in the low correlation structure and performed much worse than the SEQ-CAT item selection. Also, Subtest 1 was the first administered subtest in IND-UCAT and PC-MCAT, but not the first selected subtest in SEQ-CAT. For all of the examinees, it was the second selected subtest for administration in SEQ-CAT, which means that part of the collateral information from the first selected subtest (Subtest 3) was used for the SEQ-CAT item selection in Subtest 1. Although merely a moderate correlation (=0.40) was involved in the relevant joint marginal distribution, it contributed to the best performance of the SEQ-CAT item selection in Subtest 1 in comparison to the IND-CAT and PC-MCAT item selections.

Furthermore, another distinguishable characteristic of SEQ-CAT is the adaptive subtest selection. That is, the sequence of subtests administered in the other two methods is fixed and prespecified from Subtest 1 to Subtest 3, whereas the sequence of subtests in SEQ-CAT is adaptive to the performance of examinees in the proceeding subtests. It is worth noting that the adaptive subtest selection in SEQ-CAT is determined not only by the SEQ-CAT subtest selection algorithm, but also by the configurations of subpools, of which the latter became the primary determinant in the study.

As depicted in Chapter 3, three subpools from an operational testing program were adopted in the study. Among the three subpools, Subpool 3 had the largest number of items (320 items) and then Subpool 1 had the second largest (281 items) whereas Subpool 2 had only 154 items. Other than the unbalanced number of items in each subpool, the three IRT item parameters individually demonstrated different distributions among the three subpools, which are presented in Table 5 and Figure 3. As a consequence, the test information functions for each subpool were remarkably distinct, as is shown in Figure 10 below.

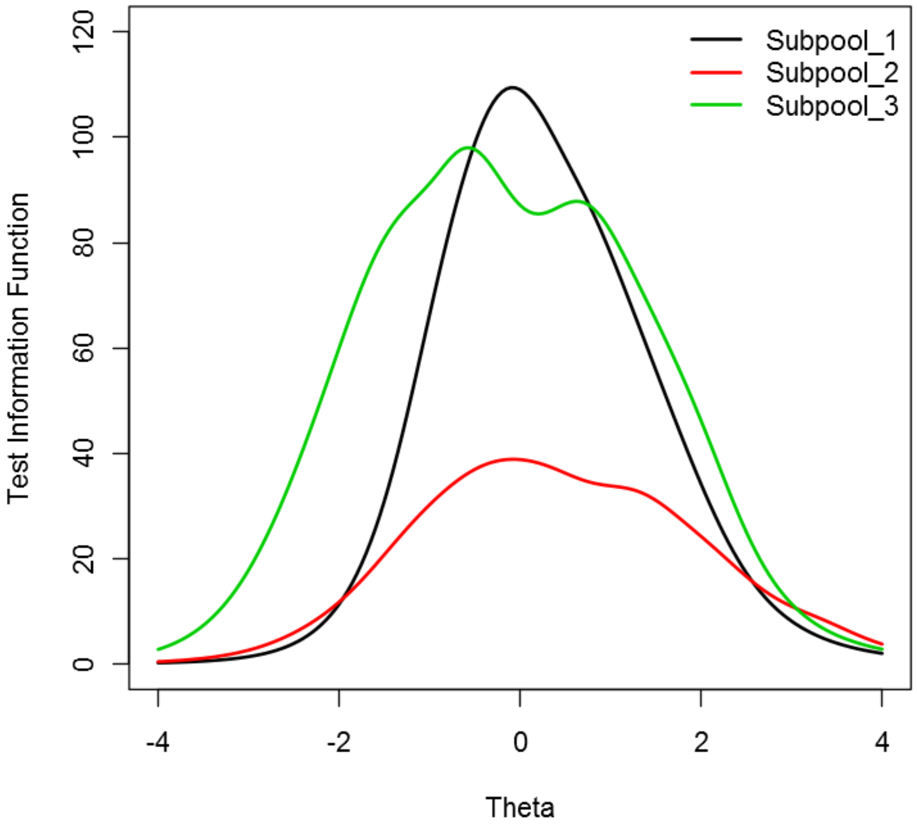

Figure 10. Test Information Function for Each Subpool.

Figure 10 shows that Subtest 3 could provide the largest amount of test information across nearly the entire ability scale, whereas the largest amount of test information provided by Subtest 1 primarily concentrated on the medium-level abilities. Comparatively, the test information provided by Subtest 2 was much lower than the other two subtests across the entire ability scale, aside from the small areas around the two ends. Because of this fact, the sequence of subtests selected by SEQ-CAT was almost the same for all the examinees, which was Subtest 3, Subtest 1, and Subtest 2. The only exception occurred in the high correlation structure, in which a few examinees (around 2.4% of examinees) had Subtest 2 as the second administered subtest in SEQ-CAT.

As indicated in Chapter 2, the subtest selection adaptation in SEQ-CAT can customize a test battery corresponding to the performance of individual examinees in the previous subtests, and may therefore improve the subscore estimates by optimizing the subtest assembly. However, the uniform sequence of subtests in SEQ-CAT of this study partly constrained the effects of the particular adaptation of SEQ-CAT on subscore estimation. That is, the configurations of the subpools from the operational testing program did not give lots of play for SEQ-CAT to adaptively select the subtests. On the other hand, the uniform sequence of subtests in SEQ-CAT still reflected the adaptation of SEQ-CAT in subtest selection because it was different from the fixed sequence in IND-UCAT and PC-MCAT. It was still determined by the criterion that the subtest providing the maximum sum of the information to the current ability estimate of an examinee should be selected. As a consequence, this uniform, but adaptive, subtest sequence provided the SEQ-CAT item selection the possibility of performing better than the PC-MCAT item selection in the mixed correlation structure.

Specifically speaking, Subpool 3 was relatively well-constructed, and therefore was selected as the first administered subtest for all the examinees in SEQ-CAT. Although there was no collateral information available for the SEQ-CAT item selection in Subtest 3, the well-constructed subpool still provided more appropriate items measuring the current ability estimate, which resulted in nearly no differences among the three item selections in Subtest 3, particularly for the correlation-based subscoring methods. As the test proceeded to administer the second subtest (Subtest 1), the sensitivity of the SEQ-CAT item selection to the moderate correlation between Subtest 1 and Subtest 3 facilitated its subscore estimation in Subtest 1. This facilitation was manifested by the large discrepancies on the three outcome measures between the SEQ-CAT and PC-MCAT item selections. In the last administered subtest (Subtest 2), the utilization of the entire correlation structure in the SEQ-CAT item selection made it competitive enough to outperform the PC-MCAT item selection. Therefore, the SEQ-CAT item selection achieved the overall best performance on subscore estimation in the mixed correlation structure.

In addition, other than the three item selection algorithms, the differences between PC-MCAT and the conventional MCAT were also investigated, and the results are shown in APPENDIX E. Figures 17 to 19 of APPENDIX E show that both methods exhibited very homogeneous trends regarding correlation and RMSE. The differences regarding bias appeared to be relatively large in short sublength. However, as the sublength increased, the differences became negligible. Theoretically, due to the different sizes of the pool(s) they adopted, the differences between PC-MCAT and MCAT were totally attributed to the different series of items selected by both methods. The simple structure in MCAT does not imply that the items providing the maximum information in Equation (59) will consecutively selected from the same subpool as they will in PC-MCAT. Once an item is selected from another subpool, it will update the entire MCAT provisional subscale estimate vector by a different amount, which departs the item selection process towards a different direction from PC-MCAT.

In this study, at least 80% of the items selected by both methods (shown in Table 26) were identical, but might not be in the same sequence in the crossed conditions of three correlation structures and two sublengths. The remarkable homogeneity of the selected items by both methods most likely led their performances on subscore estimation to the general homogenous results. Therefore, it might conclude that the constrained use of pools is of less importance to a simple-structure CAT test battery. However, as mentioned in the PC-MCAT section of Chapter 3, Kroehne, Goldhammer, and Partchev (2014) arrived at a different conclusion in some of their conditions. They indicated that more systematic investigations were required in the future for examining the effects of various configurations of item pools and correlation structures on the performances of both methods. They also pointed out that the performance of PC-MCAT was sequence-dependent and could be comparable to the performance of the conventional MCAT if the optimal sequence of subtests was identified. Given the results of PC-MCAT in this study, the sequence of subtest administration predetermined in the study might be coincidentally the optimal sequence, which probably led the performance of PC-MCAT to be comparable to MCAT. Other than the sequence of subtests, some other factors, such as the distinct configurations of subpools, could also contribute to the comparable performances between PC-MCAT and MCAT.

Given the discussion above, the optimal sequence appears to be crucial to the performance of PC-MCAT. With reference to the adaptation of subtest selection in SEQ-CAT, it is very feasible to adaptively determine the optimal sequence of subtests for each examinee in PC-MCAT, which was also suggested by Kroehne et al. in their study (2014).

In other words, the subtest to be administered in PC-MCAT can be adaptively selected for each examinee corresponding to his/her performance in the previously selected subtest(s). The PC-MCAT with adaptively sequencing subtests is called SEQ-MCAT in this study, in order to be differentiated from PC-MCAT. More precisely, as conducted in SEQ-CAT, when the first subtest is to be selected, all the subpools will be screened by comparing shadow tests for the one that can maximize the sum of the determinants of Fisher's posterior information matrix over the intended sublength. The objective function of the shadow test for SEQ-MCAT in subpool *d* can be obtained by substituting Equation (88) for Equation (40),

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[\det(\boldsymbol{I}(\hat{\boldsymbol{\theta}}, u_{n_{(d)}}) + \boldsymbol{\Phi}^{-1}) \right] x_{n_{(d)}} \ . \tag{88}$$

As the first subpool d^1 is identified, the item that maximizes the determinant of Fisher's posterior information matrix in d^1 is selected to be the first item for administration. After that, the scoring procedure, described in the MAP section of Chapter 2, and the adaptive item selection procedure, described in the PC-MCAT section of Chapter 3, are routinely conducted in d^1. Once the fixed sublength of d^1 is reached, the second subtest is to be selected from the rest of subpools by the following objective function

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[\det(\boldsymbol{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{J_{(d^1)}}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}^{J_{(d^1)}}, u_{n_{(d)}}) + \boldsymbol{\Phi}^{-1}) \right] x_{n_{(d)}} , \tag{89}$$

which is substituted for Equation (88). In Equation (89), $d \neq d^1$, and $\hat{\boldsymbol{\theta}}^{J_{(d^1)}}$ refers to the vector of the provisional subscale parameter estimates obtained after the first selected subtest is completed. Following the same logic, the test proceeds until all the subpools are selected. Once the last subtest is completed, the MAP subscores of all the subtests can be simultaneously estimated from the final updated posterior distribution by Equations (8) to (10).

As a matter of fact, the subtest and item selection procedure discussed above for SEQ-MCAT can be further simplified. For a simple-structure test battery, the magnitude of the determinant of Fisher's posterior information matrix is totally determined by the changes of the diagonal elements, each denoted as

$$I(\theta_{i(d)}, \hat{\theta}_{i(d)}) + \boldsymbol{\Phi}^{-1}[d,d]. \tag{90}$$

where $\boldsymbol{\Phi}^{-1}[d,d]$ represents the *d*th diagonal element in the inverse of the prior covariance matrix. $\boldsymbol{\Phi}^{-1}[d,d]$ is always constant for all the items in subpool *d*, and thus the first term in Equation (90) determines the change of each diagonal element as an item is added in the test. Since the subtest and item selection procedure in SEQ-MCAT always concentrates on one subpool, it implies that only one element on the diagonal is changed every time a subpool or an item is selected. The largest change on that element at one time implies that the largest determinant of the matrix is obtained. That is, to seek the item(s) that maximize(s) the change of $I(\theta_{i(d)}, \hat{\theta}_{i(d)})$ is the purpose of the objective functions (88) and (89).

As a consequence, the adaptive selection of the first subtest represented by Equation (88) for SEQ-MCAT can be simplified as

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[I(\hat{\theta}_{i(d)}, u_{n_{(d)}}) \right] x_{n_{(d)}} . \tag{91}$$

The adaptive selection for the second subtest represented by Equation (89) for SEQ-MCAT is simplified as

$$\text{maximize} \sum_{n_{(d)}=1}^{N_{(d)}} \left[I(\hat{\theta}_{i(d)}^{J_{(d^1)}}, u_{n_{(d)}}) \right] x_{n_{(d)}} , \tag{92}$$

where $\hat{\theta}_{i(d)}^{J_{(d^1)}}$ refers to the provisional subscale parameter estimate for subtest *d*, which is one of the estimates obtained after the first selected subtest is completed. According to the same logic, Equation (91) is also the simplified Equation (66) for PC-MCAT and SEQ-MCAT to adaptively select the item in subpool *d*. In fact, this simplification is likewise applicable to the item selection procedure in the conventional simple-structure MCAT if D-optimality or a Bayesian version of D-optimality is adopted. That is, the entire pool (the combination of all subpools) in MCAT is screened for the *k*th item providing the largest amount of information evaluated at the provisional subscore estimate $\hat{\theta}_{i(d)}^{k-1}$ if the item is selected from subpool *d*.

The discussion above reveals at least the following three facts for a simple-structure test battery. First, D-optimality and the Bayesian version of D-optimality conducted in a variety of MCAT tests (MCAT, PC-MCAT, and SEQ-MCAT) can be largely simplified, from operating the posterior information matrices to seeking an item with maximum information in a unidimensional space (either three subpools separately or three subpools consecutively). Second, the simplified item selection criteria can avoid the deadlock of the item selection that arises when D-optimality is employed to select the first few (three or more in this study) items in a variety of MCAT tests. Third, the Bayesian version of D-optimality for these three types of MCATs actually does not capitalize on any collateral information during the item and subtest selection procedure. Although the inverse of the covariance matrix is one of the components in the posterior information matrix, it does not play any role on adaptively selecting an item and a subtest for a simple-structure test battery.

Consequently, the finding in terms of the good performance of the PC-MCAT item selection in the high correlation structure was totally attributed to the MIRT MAP scoring procedure, as opposed to the allegation of use of collateral information in the PC-MCAT item selection procedure. More precisely, by adding the collateral information via the prior distribution, the MIRT MAP scoring procedure in the high correlation structure could most efficiently approach the true ability parameters among all the scoring methods. Correspondingly, a series of items that most approximately and optimally measured the true ability parameters were sorted out by PC-MCAT from the subpools.

In addition, regarding the conditions involving the IND-CAT and SEQ-CAT item selections in the study, there were no differences between the PC-MCAT and MCAT scoring because both of the scoring methods adopt the MIRT MAP scoring algorithms, and both of the item selection algorithms concentrate the item selection in one subpool until the subscores in that subpool are all obtained. Also, other than the original PC-MCAT and MCAT compared in the study, the other two original CAT scoring methods (IND-UCAT and SEQ-CAT) are more often investigated in the CAT studies. For future handy references, the results of these four original methods are presented in the same plots and tables in APPENDIX F, so that the discrepancies among these four original methods on score estimation can be more straightforwardly demonstrated.

Significance of the Study

As conducted in the P&P tests in the literature, the study is dedicated to examine how the five CAT subscoring methods perform in the CAT testing environment as subtest lengths and the correlations between subtests are varied. Also, to ensure the high comparability among the five subscoring methods, the distinctions on the item selection algorithms are considered. By making comparisons in this study, the advantages and disadvantages of the five subscoring methods are demonstrated and generalized under varied testing conditions. From the statistical standpoint, the differences may not be very momentous. Nevertheless, some systematical guidelines relevant to practice can still be provided for their future applications, especially in empirical studies and operational CAT testing programs. In the literature, there are no comprehensive and thorough comparison studies for these CAT subscoring methods, and therefore this study would contribute valuable sources to the literature for the interested audience.

In the study, two post-hoc subscore estimation methods, AUG-CAT and reSEQ-CAT, are also investigated along with the other three CAT-based subscoring methods. Their application to the CAT framework greatly enriches the CAT subscoring mechanism and may make it possible to implement the subscoring procedure in operational CAT tests with more ease. Relatively speaking, AUG-CAT is easy to compute and reSEQ-CAT is ideal for accurate subscores. Both methods are applicable not only in CAT tests but also in P&P tests, because they are always implemented after a conventional test regardless of the testing formats. This flexibility, on the other hand, allows for a new insight into the implementation of subscoring in a test, which is to consider the feasibility of some augmentation techniques after a conventional CAT test is administered. One of the benefits of the post-hoc augmentation is that the quality of the subscore estimates is guaranteed under the condition that the traditional unsophisticated CAT test is not interfered.

Also, as mentioned above, three item selection algorithms are separately implemented by being paired with the individual subscoring methods. On the one hand, it ensures that these subscoring methods are compared under more comparable conditions. On the other hand, it gives more possibilities of improving the subscore estimates, depending on how the collateral information is added into the item selection procedure, besides simply developing more efficient subscoring methods. These three item selection algorithms either ignore the collateral information or manifest the approach of capitalizing on the collateral information during the item selection procedure. Through the above-noted comparisons, the item selection methods that achieve the largest improvement on score estimates can be considered for future applications by being paired with the corresponding subscoring method.

Moreover, the study adopts PC-MCAT instead of MCAT as one of the subscoring methods for comparison. It is a trade-off between avoidance of item context effects and constrained use of item pools. It is also, in some sense, more applicable and comparable to apply the other subscoring methods on the items selected by PC-MCAT instead of by MCAT. However, although the performance of PC-MCAT and MCAT are comparable in this study, this compromise may most likely make the performance of PC-MCAT inferior to MCAT in some other conditions. To compensate for the negative impacts of the constrained use of item pools, this study proposes the PC-MCAT with adaptively sequencing subtests (SEQ-MCAT) for future investigations. In the meantime, the simplified item selection criteria in a simple-structure MCAT, PC-MCAT, and SEQ-MCAT are suggested by the study. The simplifications can not only avoid the deadlock of the indefinable ability estimates in MCAT mentioned by Segall (1996), but also facilitate the applications of MCAT, PC-MCAT, and SEQ-MCAT in practice.

Last but not least, based on the hierarchical latent trait structure, the successive scoring procedure suggested in the study could provide interested parties with both subscores and total scores from one test, and thus achieve the testing purposes of ranking and diagnosis at the same time. This procedure is easy to conduct, and the guidelines of its use are also given in the study in order to help determine under what conditions and how this procedure could be applied. In addition, to better fit the requirements of subscoring, this successive scoring procedure may provide a new clue and possibility of adjusting the item calibration system. For example, when Wainer's AUG (2001) was used to estimate subscores in studies and operational P&P tests, the bank of item parameters was actually established from test-based item calibration instead of subtest-based item calibration, which means that all the items in each subtest were assembled and calibrated as a whole, as also implemented in van der Linden's (2010) study for SEQ-CAT. The item parameters calibrated in this way are more suitable to estimate total scores rather than subscores. Otherwise, the subtest-based item calibration is required. Currently, according to the successive scoring procedure, if the item bank is calibrated based on the higher-order IRT model, the item parameters are appropriate to the estimation of both total scores and subscores.

Limitations and Future Directions

Given the simulation design and the corresponding results, there are a few limitations in the study. First, this study investigated a test battery with only three subtests. The small number of subtests may largely constrict the performances of the subscoring and item selection methods. For example, as mentioned previously, the SEQ-CAT subscoring and item selection methods take less advantage of the collateral information for the first few subtests compared to the other correlation-based methods. The results in Chapter 4 imply that this characteristic of SEQ-CAT may either favor or impair its performances depending on the correlations among subtests. However, if the number of subtests increases, it is open to doubt that whether the improvement or the impairment attributed to this characteristic is still validated by the change of the correlation structures. It is also called in question that the discrepancy between SEQ-CAT and the other methods resulted from this characteristic tends to be even larger or smaller. For the total score estimation, if more subtests are included in a test, each subscore will become less influential to the total score estimate, and the negative impact of this characteristic in SEQ-CAT may also become insignificant for the high correlation structure.

Other than the impact on SEQ-CAT, the number of subtests also has a strong impact on the total score estimation for all the other subscoring methods. Based on the properties of the likelihood function, the more subtests included in a test battery, the more accurate the total score estimate will be. Correspondingly, if there are sufficient subtests in the low correlation structure, more subscores used to estimate the total score may compensate for less collateral information being accessible to total score estimation. Therefore, the total score estimation procedure suggested by the study may be reconsidered to apply for the conditions of the low correlation structure. In addition, as depicted for the high correlation structure, the total score estimation of all the methods demonstrated a better performance than their subscore estimation. There is some possibility that the increase of the number of subtests may further enhance their performance on total score estimation in the high correlation structure, and therefore guarantee the quality of total scores for test users to make high-stake decisions.

Also, the study employed three subpools from an operational testing program and to some extent, took account of the reflection of the real testing realm. However, these three subpools were not originally constructed for implementing the subscoring procedure as a whole. As described in Chapter 3, the number of items and the distributions of item parameters were considerably different among the three subpools, so that the performances of the subscoring methods and the item and subtest selection methods heavily depended on the nature of the subpools. For instance, all of the subscoring methods exhibited the best performance in Subtest 3 in all the conditions because of its superior construction. This impact was particularly critical to SEQ-CAT, which was demonstrated by the adaptive, but uniform sequence of subtest selection for all the examinees in almost all the correlation structures. This adaptive sequence was primarily determined by the large discrepancy on the properties of subpools, instead of on the performances of individual examinees in the previous selected subtests. The restricted subtest selection in SEQ-CAT correspondingly influenced the performances of the SEQ-CAT subscoring and item selection algorithms. Aside from SEQ-CAT, the impact was relatively less crucial to the other subscoring and item selection methods in the study, because their performances on score estimation were not closely associated to the sequence of subtest administration, and also all of them were compared primarily within each subtest rather than between subtests.

On the other hand, the sequence of subtest administration in PC-MCAT is predetermined and fixed, which was identical to the sequence in IND-UCAT of the study. Although PC-MCAT competed with MCAT in the study, the sequence of administering subtests is influential to the performance of PC-MCAT based on the findings of Kroehne, Goldhammer, and Partchev’s study (2014). If the optimal sequence of subtest administration can be identified for PC-MCAT, the constrained use of item pools in PC-MCAT might be largely compensated for. Therefore, SEQ-MCAT proposed in the study is worth investigating, in which the sequence of subtest administration is adaptively searched for PC-MCAT.

Besides, in the mixed correlation structure, the high correlation emerged between Subtest 2 and Subtest 3, which was advantageous to the performance of the PC-MCAT item selection. However, the good construction of Subpool 3 partly shrunk the distinctions between the PC-MCAT item selection and the other two item selections in Subtest 3. On the other hand, the PC-MCAT item selection was insensitive to the moderate correlations between Subtest 1 and the other two subtests. Therefore, the overall performance of the PC-MCAT item selection on subscore estimation might be attenuated because of this pattern of correlations in the mixed correlation structure. If the high correlation emerges between Subtest 1 and Subtest 2, it will be in question that the SEQ-CAT item selection outperforms the PC-MCAT item selection in the mixed correlation structure. Also, there are many other possibilities regarding the configurations of subpools in practice, which may provide different patterns of the performances of these subscoring and item selection methods. For example, the same number of items is included in some subpools, of which the maximum test information functions center at different ability levels.

Furthermore, as illustrated previously, three item selection algorithms were separately paired with each subscoring method in order to fulfill high comparability. That is, the original IND-CAT, SEQ-CAT, and PC-MCAT were individually implemented, and then all the other four subscoring methods were applied to the items selected by these three methods. In this way, all the subscoring methods were compared based on the same collection of items. Due to the purpose of comparison, this is the defined combination of the subscoring methods and item selection algorithms in the study, which is actually a post-hoc combination. As a matter of fact, this attempt, on the other hand, triggers another way of thinking, which could be perceived as the just-in-time combination of the subscoring method and the item selection algorithm. More precisely, the combination is to use one subscoring method (e.g. PC-MCAT) to obtain the real-time score estimate, and then to use another item selection algorithm (e.g. SEQ-CAT) to select the most appropriate item measuring that real-time score estimate.

Additionally, three different levels of the correlation structures were considered in the study, which cannot fully represent the correlation patterns among subtests in practice. The limited number of subtests especially provides few possibilities to demonstrate more patterns of correlations, such as the mixture of low and moderate correlations or the mixture of three levels of correlations in one correlation structure. In the study, the collateral information exploited for score estimation mainly referred to the correlation information among subtests. In fact, some other in-test and/or out-of-test collateral information, such as some demographic variables, can be considered for use and comparisons. As is known, the total scores are always employed for high-stake decisions. To ensure this purpose of the scoring, it is still required to conduct substantive studies relevant to the total score estimation approach suggested in the study. Besides, more item formats can be considered to apply in conjunction with these subscoring and item selection methods.

To sum up, there are a few possible directions in the future: (1) to increase the number of subtests; (2) to employ the subpools with the configurations distinct from the current study; (3) to examine the performances of SEQ-MCAT proposed in this study by comparison to PC-MCAT and MCAT; (4) to investigate the feasibility and efficiency of the just-in-time combination of the subscoring method and the item selection algorithm; (5) to explore the different patterns of the correlation structures; (6) to exploit some other sources of collateral information, such as the demographic information, in the subscoring procedure for comparison; (7) to apply the subscoring and item selection methods investigated in the study to other item formats (e.g. polytomously-scored items) in a CAT test; (8) to conduct more studies on the total score estimation procedure suggested in the study.

REFERENCES

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27*, 241–253.

Allen, D.D., Ni, P.S., & Haley, S.M. (2008). Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disability and Rehabilitation*, *30*(6), 479–484.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measuremen*t, *6*, 431-444.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443-459.

Chang, H.H., & Ying, Z.L. (1996). A global information approach to computerized adoptive testing. *Applied Psychological Measurement*, *20*(3), 213–229.

Chen, P. H. (2009). Comparison of adaptive Bayesian estimation and weighted Bayesian estimation in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Cheng, Y. & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369–383.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: CBS College.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement, 32*, 355-370.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement, 34,* 267-285.

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

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

DeMars, C. E. (2005). *Scoring subscales using multidimensional item response theory models*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Deng, H., Ansley, T., & Chang, H.-H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47,* 202-226.

Dorans, N. J. (2005). *Why trait scores can be problematic.* Paper presented to the ETS Visiting Panel on Research, Princeton, NJ: Educational Testing Service.

Dwyer, A., Boughton, K. A., Yao, L., Lewis, D., & Steffen, M. (2006). *A comparison of subscore augmentation methods using empirical data.* Paper presented at the National Council on Measurement in Education (NCME), San Francisco, CA.

Fu, J., & Qu, Y. (2010). *A comparison of subscore reporting approaches in simulated data.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Denver, CO.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31,* 241-259.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145-220.

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

Huang, H-Y., Chen, P-H., & Wang, W-C. (2012). Computerized adaptive testing using a class of higher-order item response theory models. *Applied Psychological Measurement*, 36, 689-706.

Huang, H-Y., Wang, W-C., Chen, P-H., & Su, C-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*, 619-637.

Kahraman, N., & Kamata, A. (2004). Increasing the precisions of subscale scores by using out-ofscale information. *Applied Psychological Measurement, 28,* 407-426.

Kelley, T. L. (1927). The interpretation of educational measurements. New York: World Book.

Kelley, T. L. (1947). Fundamentals of statistics. Cambridge, MA: Harvard University Press.

Kingsbury, G. G. & Zara, A. R. (1991). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

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

Lee, Y-H., Ip, E.H., & Fuh, C-D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, *68(2),* 215-232.

Li, Y.H., & Schafe,W. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, *29,* 3–25.

Liu, F., Li, J., & Choi, S. (2014). *Subscoring by sequencing an adaptive testing battery*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) in Philadelphia, PA.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale: Lawrence Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Luo, X., Diao, Q., & Ren, H. (2014). *Subscale Reporting in CAT Using the Augmented Subscore Approach.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) in Philadelphia, PA.

Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*(2), 273–296.

Monaghan, W. (2006). *The fact about subscores.* (ETS RDC-04). Princeton, NJ: Educational Testing Service.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standard* . Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: The National Academies Press.

Nicewander, W. A. & Thomasson, G, L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement, 23*, 239-247.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

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

Segall, D.O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*, *66*, 79–97.

Segall, D.O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 57–75). New York, NY: Springer.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *6*, 899–919.

Shin, C. D., Ansley, T., Tsai, T., & Mao X. (2005). *A comparison of methods of estimating objective scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Shin, D. (2007). A comparison of method of estimating subscale scores for Mixed-Format tests. *Pearson Educational Measurement*.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (ETS Research Memorandum No. RR-08-18). Princeton, NJ: Educational Testing Service.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic subscores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70(3),* 357-375.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: a case study when the test is essentially unidimensional. *Applied Measeurement in Education, 23:* 63-86.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51,* 589-601.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151–166.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17(2),* 89-112.

U.S. Department of Education. (2002). *No Child Left Behind Act of 2001* (Pub. L. No. 107–110, §1111, 115 STAT. 1449-1452). Washington, DC: U.S. Department of Education.

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216.

van der Linden, W.J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

van der Linden, W. J. (2005). Linear models for optimal test design. New York: Springer-Verlag.

van der Linden, W.J. (2010). Constrained Adaptive Testing with Shadow Tests. *Elements of Adaptive Test*(pp.31-55). New York: Springer.

van der Linden, W.J. (2010). Sequencing an Adaptive Test Battery. *Elements of Adaptive Test*(pp.103-119). New York: Springer.

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575–588.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika*, *54*, 427–450.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement, 37*(2), 113-140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B., Rosa, K., Nelson, L., & Swygert, K. A. (2001). Augmented scores: ‘‘Borrowing strength’’ to compute scores based on small number of items. In D. Thissen & H. Wainer(Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83–105.

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

Yen, W. M., Sykes, R. C., Ito, K., & Julian, M. (1997). *A Bayesian/IRT index of objective performance for tests with mixed item types.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

APPENDIX A

TABLES OF THE CORRELATION (ORIGINAL VALUES) BETWEEN θ AND $\hat{\theta}$

Table 17

Correlation (Original Values) between θ and $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.698	0.698	0.698	0.698	0.698	0.711	0.711	0.711	0.711	0.711
	SEQ-CAT	0.697	0.697	0.697	0.697	0.697	0.711	0.711	0.711	0.711	0.711
	PC-MCAT	0.696	0.696	0.696	0.696	0.696	0.710	0.71	0.710	0.710	0.710
Sub_COMB	IND-UCAT	0.942	0.942	0.943	0.943	0.942	0.971	0.971	0.971	0.971	0.971
	SEQ-CAT	0.941	0.942	0.942	0.942	0.942	0.970	0.971	0.971	0.971	0.971
	PC-MCAT	0.940	0.941	0.941	0.941	0.941	0.971	0.971	0.971	0.971	0.971
Subtest 1	IND-UCAT	0.942	0.942	0.943	0.943	0.942	0.971	0.971	0.971	0.971	0.971
	SEQ-CAT	0.941	0.942	0.942	0.942	0.942	0.970	0.970	0.971	0.970	0.970
	PC-MCAT	0.938	0.938	0.939	0.939	0.939	0.971	0.971	0.971	0.971	0.971
Subtest 2	IND-UCAT	0.925	0.925	0.926	0.926	0.927	0.962	0.962	0.962	0.962	0.962
	SEQ-CAT	0.924	0.925	0.925	0.925	0.925	0.962	0.962	0.962	0.962	0.962
	PC-MCAT	0.927	0.927	0.928	0.928	0.928	0.962	0.962	0.962	0.962	0.962
Subtest 3	IND-UCAT	0.958	0.959	0.959	0.959	0.959	0.979	0.980	0.980	0.980	0.979
	SEQ-CAT	0.958	0.958	0.959	0.959	0.959	0.979	0.979	0.980	0.980	0.979
	PC-MCAT	0.957	0.957	0.957	0.957	0.957	0.980	0.980	0.980	0.980	0.980

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 7, they were caused by rounding errors.

Table 18

Correlation (Original Values) between θ and $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.915	0.916	0.922	0.920	0.921	0.933	0.934	0.937	0.936	0.936
	SEQ-CAT	0.921	0.925	0.926	0.926	0.925	0.935	0.938	0.937	0.937	0.937
	PC-MCAT	0.916	0.916	0.921	0.922	0.921	0.935	0.935	0.937	0.937	0.937
Sub_COMB	IND-UCAT	0.940	0.943	0.947	0.946	0.947	0.969	0.969	0.971	0.971	0.971
	SEQ-CAT	0.945	0.948	0.951	0.951	0.950	0.970	0.971	0.972	0.972	0.972
	PC-MCAT	0.942	0.943	0.947	0.947	0.946	0.970	0.970	0.971	0.971	0.971
Subtest 1	IND-UCAT	0.941	0.941	0.943	0.944	0.943	0.969	0.969	0.970	0.970	0.970
	SEQ-CAT	0.946	0.946	0.947	0.947	0.947	0.970	0.971	0.971	0.971	0.971
	PC-MCAT	0.934	0.934	0.937	0.936	0.936	0.968	0.968	0.969	0.969	0.969
Subtest 2	IND-UCAT	0.929	0.932	0.941	0.939	0.940	0.960	0.961	0.965	0.965	0.965
	SEQ-CAT	0.939	0.948	0.948	0.948	0.948	0.961	0.965	0.965	0.965	0.965
	PC-MCAT	0.932	0.934	0.943	0.943	0.942	0.961	0.962	0.965	0.965	0.965
Subtest 3	IND-UCAT	0.951	0.956	0.957	0.956	0.956	0.977	0.979	0.979	0.979	0.978
	SEQ-CAT	0.951	0.951	0.957	0.957	0.957	0.977	0.977	0.979	0.979	0.978
	PC-MCAT	0.959	0.961	0.962	0.961	0.962	0.979	0.980	0.980	0.980	0.980

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 8, they were caused by rounding errors.

Table 19

Correlation (Original Values) between θ and $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.961	0.965	0.966	0.966	0.963	0.976	0.976	0.977	0.977	0.976
	SEQ-CAT	0.963	0.960	0.968	0.968	0.965	0.976	0.976	0.977	0.977	0.977
	PC-MCAT	0.966	0.966	0.968	0.969	0.966	0.977	0.976	0.978	0.978	0.977
Sub_COMB	IND-UCAT	0.937	0.952	0.962	0.961	0.957	0.967	0.972	0.977	0.977	0.976
	SEQ-CAT	0.946	0.957	0.965	0.965	0.962	0.970	0.976	0.978	0.977	0.977
	PC-MCAT	0.944	0.952	0.965	0.965	0.961	0.968	0.971	0.977	0.977	0.976
Subtest 1	IND-UCAT	0.946	0.946	0.963	0.963	0.961	0.969	0.969	0.978	0.978	0.976
	SEQ-CAT	0.955	0.964	0.967	0.967	0.965	0.971	0.978	0.979	0.979	0.977
	PC-MCAT	0.940	0.940	0.966	0.966	0.959	0.966	0.966	0.977	0.977	0.974
Subtest 2	IND-UCAT	0.916	0.941	0.953	0.953	0.948	0.957	0.965	0.971	0.970	0.970
	SEQ-CAT	0.935	0.957	0.958	0.958	0.956	0.961	0.972	0.972	0.972	0.971
	PC-MCAT	0.929	0.943	0.956	0.956	0.955	0.958	0.965	0.971	0.971	0.970
Subtest 3	IND-UCAT	0.950	0.969	0.969	0.969	0.965	0.977	0.982	0.982	0.982	0.982
	SEQ-CAT	0.950	0.950	0.970	0.970	0.966	0.977	0.977	0.982	0.982	0.982
	PC-MCAT	0.964	0.973	0.973	0.973	0.971	0.980	0.983	0.983	0.983	0.983

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 9, they were caused by rounding errors.

APPENDIX B

TABLES OF THE BIAS (ORIGINAL VALUES) OF $\hat{\theta}$

Table 20

Bias (Original Values) of $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.034	0.034	0.034	0.033	0.037	0.021	0.021	0.021	0.020	0.021
	SEQ-CAT	0.037	0.037	0.037	0.036	0.039	0.021	0.020	0.021	0.020	0.021
	PC-MCAT	0.046	0.046	0.047	0.045	0.050	0.024	0.024	0.024	0.024	0.024
Sub_COMB	IND-UCAT	0.024	0.023	0.024	0.022	0.026	0.010	0.011	0.011	0.009	0.011
	SEQ-CAT	0.026	0.026	0.027	0.026	0.029	0.010	0.010	0.010	0.010	0.011
	PC-MCAT	0.036	0.036	0.036	0.035	0.040	0.013	0.014	0.014	0.013	0.014
Subtest 1	IND-UCAT	0.020	0.020	0.022	0.020	0.026	0.011	0.011	0.010	0.009	0.013
	SEQ-CAT	0.023	0.024	0.025	0.024	0.029	0.010	0.010	0.010	0.009	0.013
	PC-MCAT	0.035	0.035	0.035	0.032	0.042	0.014	0.014	0.014	0.014	0.017
Subtest 2	IND-UCAT	0.038	0.038	0.038	0.036	0.041	0.018	0.020	0.019	0.018	0.019
	SEQ-CAT	0.043	0.042	0.044	0.042	0.047	0.019	0.018	0.019	0.018	0.019
	PC-MCAT	0.054	0.054	0.054	0.052	0.060	0.022	0.024	0.023	0.023	0.023
Subtest 3	IND-UCAT	0.013	0.011	0.012	0.011	0.012	0.002	0.002	0.003	0.002	0.001
	SEQ-CAT	0.013	0.013	0.012	0.011	0.012	0.002	0.002	0.003	0.002	0.001
	PC-MCAT	0.019	0.019	0.021	0.019	0.018	0.003	0.003	0.004	0.003	0.002

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 10, they were caused by rounding errors.

Table 21

Bias (Original Values) of $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.057	0.055	0.057	0.050	0.062	0.032	0.032	0.034	0.032	0.033
	SEQ-CAT	0.046	0.048	0.051	0.047	0.050	0.026	0.027	0.028	0.027	0.027
	PC-MCAT	0.052	0.052	0.055	0.050	0.056	0.032	0.033	0.034	0.032	0.033
Sub_COMB	IND-UCAT	0.039	0.036	0.040	0.035	0.044	0.018	0.018	0.019	0.018	0.020
	SEQ-CAT	0.037	0.038	0.039	0.035	0.041	0.017	0.017	0.017	0.016	0.018
	PC-MCAT	0.041	0.040	0.043	0.039	0.046	0.022	0.022	0.022	0.021	0.023
Subtest 1	IND-UCAT	0.046	0.046	0.048	0.047	0.051	0.029	0.029	0.030	0.030	0.032
	SEQ-CAT	0.053	0.052	0.053	0.051	0.059	0.032	0.030	0.031	0.030	0.034
	PC-MCAT	0.059	0.059	0.059	0.058	0.066	0.039	0.039	0.038	0.038	0.042
Subtest 2	IND-UCAT	0.044	0.044	0.045	0.038	0.049	0.020	0.020	0.022	0.019	0.020
	SEQ-CAT	0.031	0.033	0.038	0.033	0.034	0.012	0.013	0.015	0.013	0.012
	PC-MCAT	0.038	0.039	0.042	0.036	0.042	0.019	0.020	0.021	0.018	0.019
Subtest 3	IND-UCAT	0.027	0.020	0.026	0.020	0.032	0.006	0.004	0.005	0.004	0.008
	SEQ-CAT	0.027	0.027	0.025	0.021	0.032	0.006	0.006	0.005	0.004	0.008
	PC-MCAT	0.027	0.023	0.027	0.023	0.031	0.007	0.006	0.007	0.006	0.008

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 11, they were caused by rounding errors.

Table 22

Bias (Original Values) of $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.023	0.015	0.015	0.007	0.024	0.007	0.005	0.004	0.001	0.006
	SEQ-CAT	0.015	0.006	0.008	0.002	0.015	0.005	-0.001	0.002	0.000	0.004
	PC-MCAT	0.031	0.024	0.025	0.017	0.032	0.009	0.006	0.006	0.002	0.009
Sub_COMB	IND-UCAT	0.035	0.026	0.020	0.013	0.037	0.015	0.012	0.008	0.005	0.015
	SEQ-CAT	0.022	0.008	0.012	0.006	0.022	0.012	0.003	0.006	0.003	0.011
	PC-MCAT	0.038	0.031	0.027	0.020	0.040	0.014	0.011	0.008	0.005	0.014
Subtest 1	IND-UCAT	0.036	0.036	0.023	0.018	0.038	0.020	0.020	0.011	0.008	0.020
	SEQ-CAT	0.023	0.011	0.015	0.011	0.023	0.016	0.006	0.009	0.006	0.016
	PC-MCAT	0.045	0.045	0.029	0.022	0.048	0.020	0.020	0.011	0.009	0.020
Subtest 2	IND-UCAT	0.063	0.039	0.026	0.019	0.070	0.028	0.018	0.013	0.009	0.029
	SEQ-CAT	0.036	0.009	0.015	0.009	0.040	0.022	0.006	0.009	0.006	0.022
	PC-MCAT	0.046	0.033	0.028	0.021	0.051	0.019	0.011	0.008	0.004	0.019
Subtest 3	IND-UCAT	0.005	0.002	0.010	0.002	0.004	-0.003	-0.002	0.001	-0.002	-0.004
	SEQ-CAT	0.005	0.005	0.006	0.000	0.004	-0.003	-0.003	0.000	-0.003	-0.004
	PC-MCAT	0.022	0.017	0.024	0.017	0.022	0.005	0.002	0.006	0.002	0.004

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 12, they were caused by rounding errors.

APPENDIX C

TABLES OF THE RMSE (ORIGINAL VALUES) OF $\hat{\theta}$

Table 23

RMSE (Original Values) of $\hat{\theta}$ for All Conditions with A Low Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.729	0.728	0.728	0.728	0.728	0.715	0.715	0.715	0.715	0.714
	SEQ-CAT	0.730	0.729	0.729	0.729	0.729	0.715	0.714	0.714	0.714	0.714
	PC-MCAT	0.732	0.732	0.731	0.731	0.731	0.716	0.716	0.716	0.716	0.716
Sub_COMB	IND-UCAT	0.341	0.340	0.338	0.338	0.339	0.244	0.243	0.242	0.242	0.243
	SEQ-CAT	0.343	0.341	0.340	0.340	0.342	0.245	0.244	0.243	0.243	0.244
	PC-MCAT	0.347	0.346	0.344	0.344	0.346	0.244	0.244	0.243	0.243	0.243
Subtest 1	IND-UCAT	0.339	0.339	0.336	0.336	0.338	0.243	0.243	0.241	0.241	0.242
	SEQ-CAT	0.341	0.340	0.339	0.339	0.341	0.246	0.244	0.244	0.244	0.245
	PC-MCAT	0.353	0.353	0.349	0.349	0.352	0.243	0.243	0.242	0.242	0.243
Subtest 2	IND-UCAT	0.388	0.387	0.384	0.385	0.384	0.279	0.277	0.277	0.277	0.277
	SEQ-CAT	0.390	0.388	0.387	0.388	0.388	0.279	0.277	0.277	0.277	0.278
	PC-MCAT	0.386	0.385	0.383	0.383	0.384	0.280	0.279	0.278	0.279	0.279
Subtest 3	IND-UCAT	0.290	0.287	0.288	0.287	0.288	0.204	0.203	0.203	0.203	0.204
	SEQ-CAT	0.290	0.290	0.288	0.288	0.288	0.204	0.204	0.203	0.203	0.204
	PC-MCAT	0.296	0.294	0.294	0.294	0.295	0.202	0.202	0.202	0.202	0.202

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 13, they were caused by rounding errors.

Table 24

RMSE (Original Values) of $\hat{\theta}$ for All Conditions with A Mixed Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.417	0.414	0.399	0.403	0.401	0.366	0.365	0.358	0.358	0.358
	SEQ-CAT	0.402	0.388	0.388	0.387	0.389	0.361	0.354	0.355	0.354	0.355
	PC-MCAT	0.414	0.412	0.399	0.397	0.401	0.362	0.362	0.356	0.356	0.356
Sub_COMB	IND-UCAT	0.354	0.346	0.335	0.337	0.337	0.257	0.254	0.246	0.246	0.248
	SEQ-CAT	0.340	0.331	0.323	0.323	0.325	0.253	0.247	0.244	0.243	0.245
	PC-MCAT	0.351	0.346	0.335	0.335	0.337	0.254	0.252	0.245	0.246	0.247
Subtest 1	IND-UCAT	0.356	0.356	0.350	0.349	0.351	0.260	0.260	0.254	0.254	0.257
	SEQ-CAT	0.345	0.342	0.340	0.339	0.342	0.254	0.251	0.250	0.249	0.251
	PC-MCAT	0.377	0.377	0.371	0.371	0.372	0.264	0.264	0.259	0.259	0.261
Subtest 2	IND-UCAT	0.384	0.376	0.351	0.356	0.354	0.288	0.285	0.270	0.270	0.271
	SEQ-CAT	0.354	0.329	0.329	0.329	0.330	0.282	0.268	0.268	0.268	0.268
	PC-MCAT	0.375	0.370	0.346	0.344	0.350	0.283	0.280	0.269	0.268	0.269
Subtest 3	IND-UCAT	0.320	0.302	0.302	0.302	0.304	0.219	0.211	0.210	0.211	0.213
	SEQ-CAT	0.320	0.320	0.300	0.300	0.303	0.219	0.219	0.210	0.209	0.213
	PC-MCAT	0.293	0.284	0.284	0.284	0.284	0.208	0.205	0.203	0.205	0.205

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 14, they were caused by rounding errors.

Table 25

RMSE (Original Values) of $\hat{\theta}$ for All Conditions with A High Correlation Structure

Test	Item Selection Method	10-Item Sublength					20-Item Sublength				
		IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT	IND-UCAT	SEQ-CAT	PC-MCAT	reSEQ-CAT	AUG-CAT
Total	IND-UCAT	0.295	0.272	0.266	0.264	0.279	0.226	0.224	0.218	0.217	0.221
	SEQ-CAT	0.286	0.287	0.258	0.257	0.268	0.224	0.224	0.216	0.216	0.219
	PC-MCAT	0.273	0.267	0.256	0.254	0.264	0.219	0.220	0.214	0.213	0.216
Sub_COMB	IND-UCAT	0.358	0.315	0.282	0.281	0.298	0.258	0.240	0.218	0.218	0.225
	SEQ-CAT	0.330	0.295	0.268	0.267	0.279	0.249	0.223	0.214	0.214	0.219
	PC-MCAT	0.338	0.314	0.270	0.268	0.284	0.256	0.242	0.217	0.217	0.223
Subtest 1	IND-UCAT	0.331	0.331	0.273	0.272	0.285	0.250	0.250	0.209	0.208	0.222
	SEQ-CAT	0.301	0.268	0.256	0.255	0.268	0.240	0.210	0.206	0.206	0.216
	PC-MCAT	0.349	0.349	0.263	0.262	0.294	0.260	0.260	0.214	0.214	0.228
Subtest 2	IND-UCAT	0.416	0.350	0.314	0.312	0.336	0.300	0.270	0.248	0.248	0.253
	SEQ-CAT	0.365	0.297	0.295	0.295	0.302	0.286	0.242	0.242	0.242	0.246
	PC-MCAT	0.384	0.344	0.302	0.300	0.310	0.296	0.270	0.247	0.247	0.250
Subtest 3	IND-UCAT	0.319	0.255	0.257	0.255	0.270	0.217	0.193	0.193	0.193	0.194
	SEQ-CAT	0.319	0.319	0.250	0.249	0.264	0.217	0.217	0.192	0.192	0.193
	PC-MCAT	0.273	0.238	0.240	0.238	0.244	0.205	0.186	0.185	0.186	0.186

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 15, they were caused by rounding errors.

APPENDIX D

FIGURES OF THE CONDITIONAL BIAS OF $\hat{\theta}$

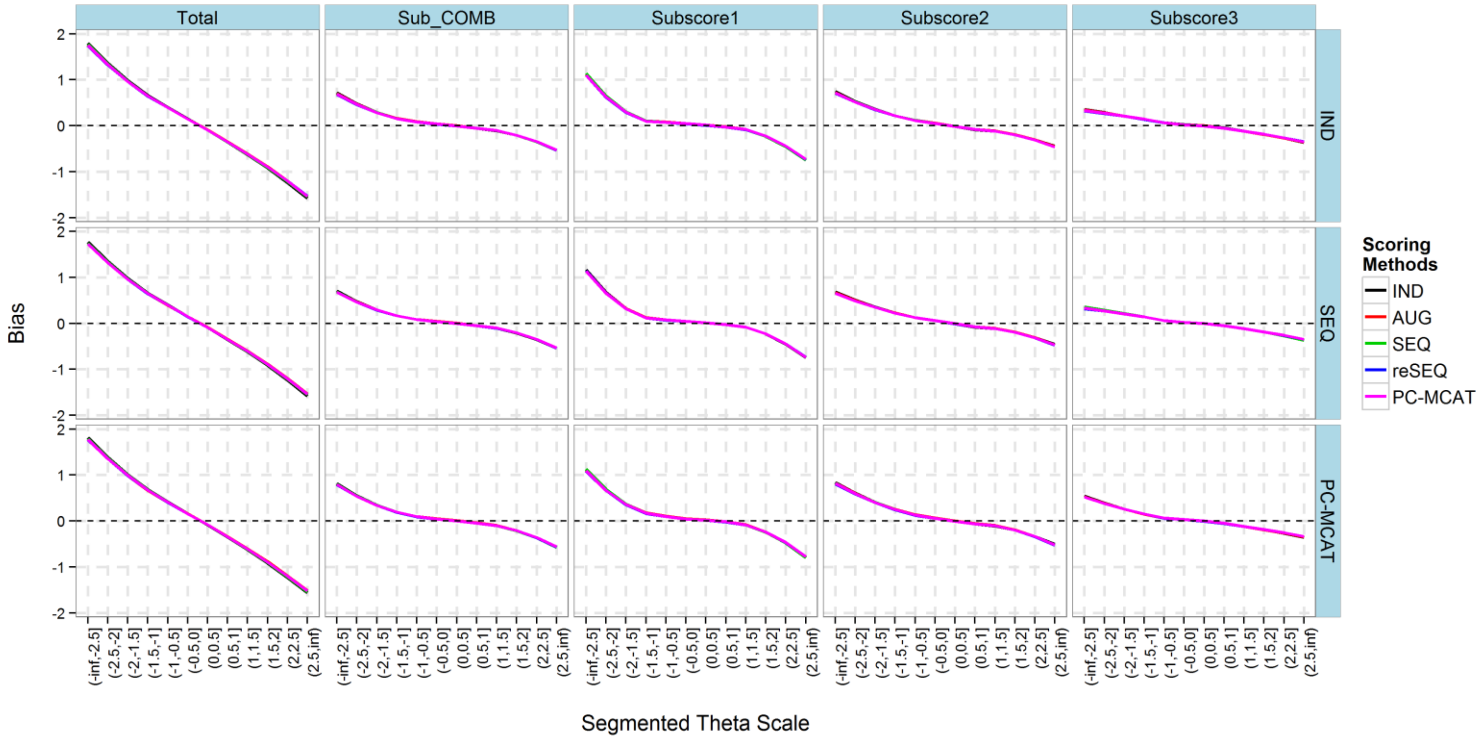

Figure 11. Conditional Bias of $\hat{\theta}$ for All the Conditions with A Low Correlation Structure and A 10-Item Sub-length.
Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

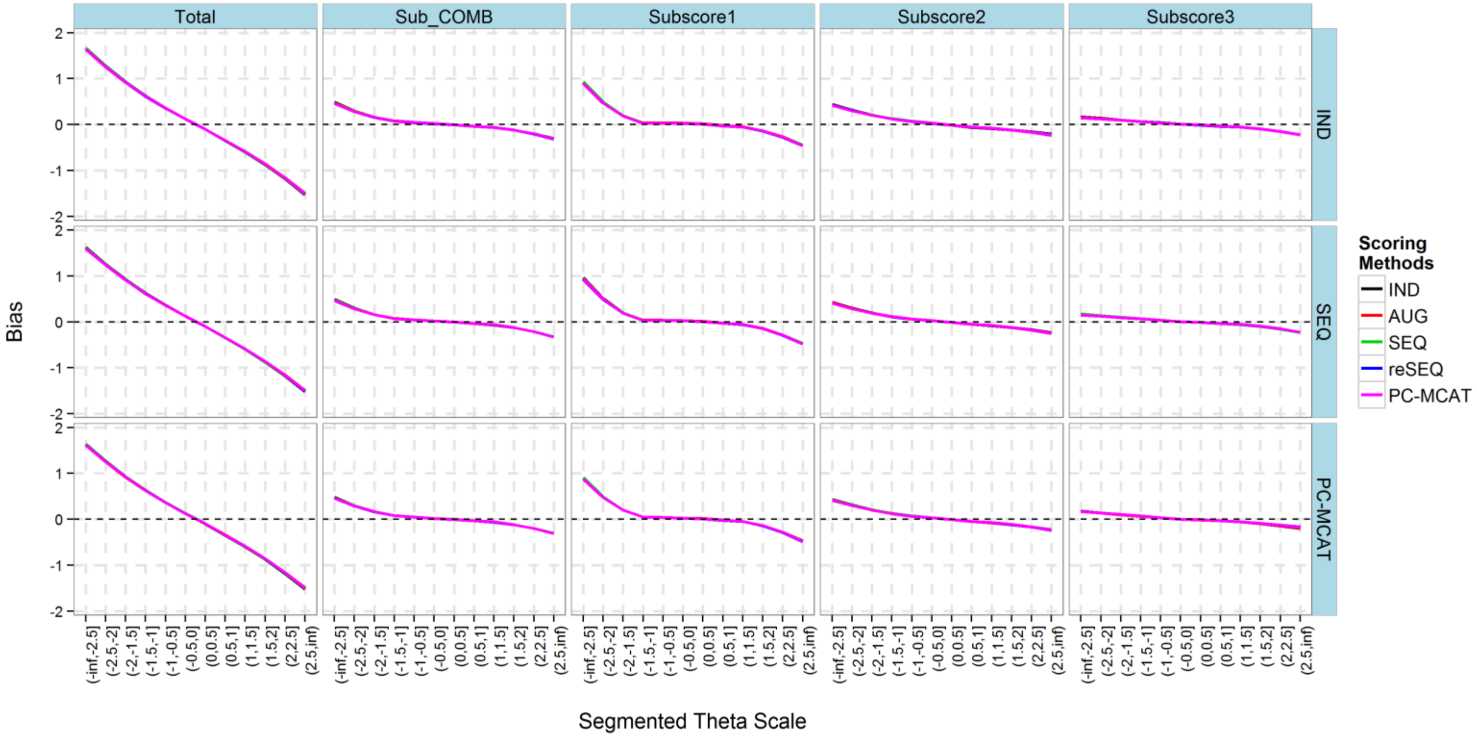

Figure 12. Conditional Bias of $\hat{\theta}$ for All the Conditions with A Low Correlation Structure and A 20-Item Sub-length.
Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

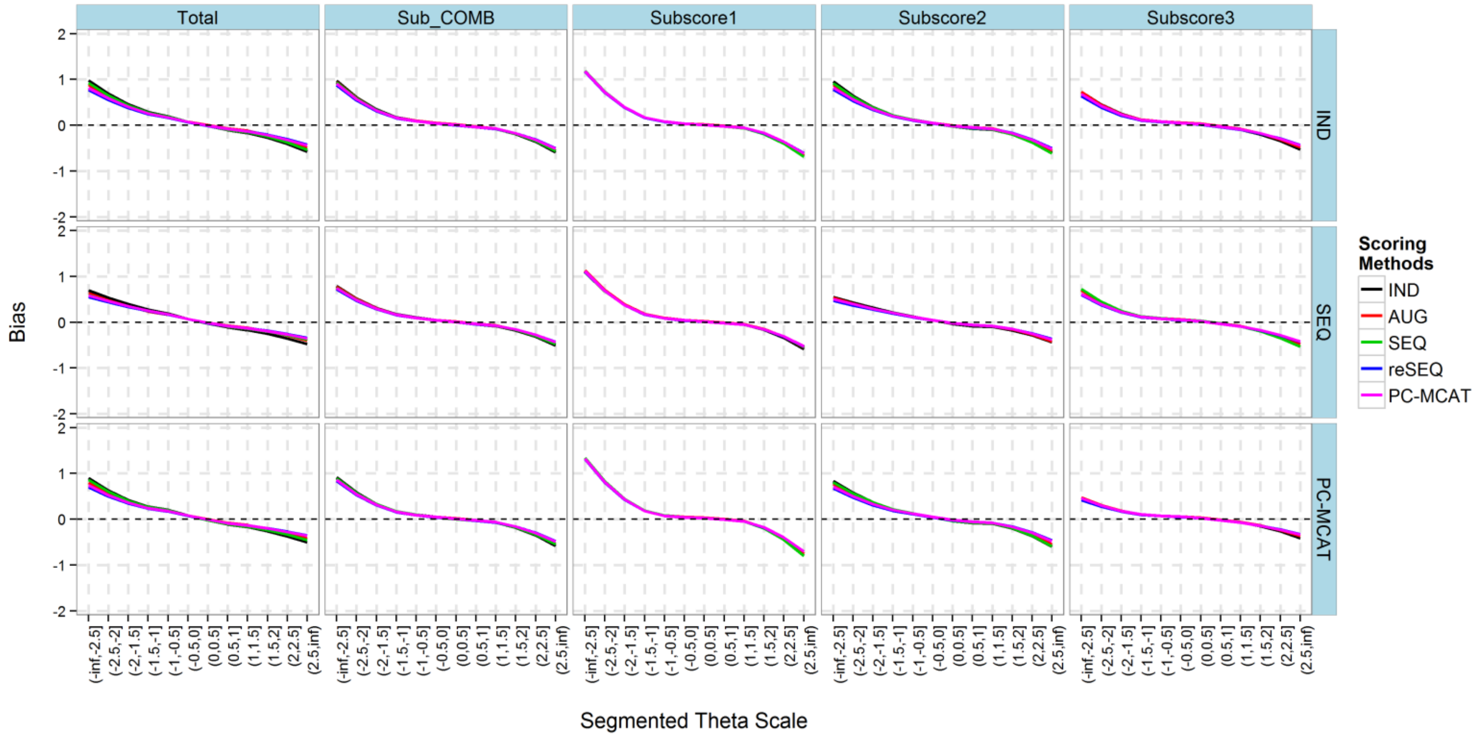

Figure 13. Conditional Bias of $\hat{\theta}$ for All the Conditions with A Mixed Correlation Structure and A 10-Item Sub-length.
Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

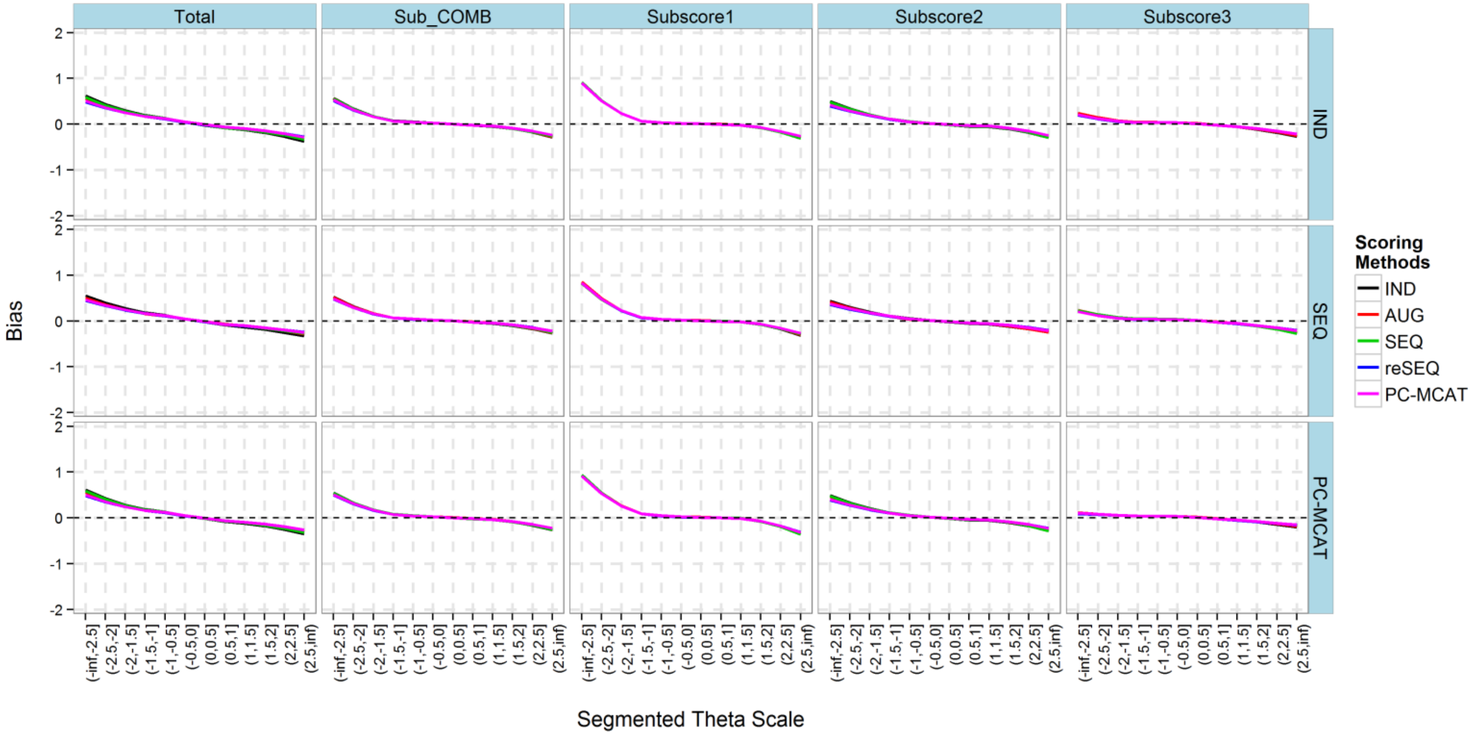

Figure 14. Conditional Bias of $\hat{\theta}$ for All the Conditions with A Mixed Correlation Structure and A 20-Item Sub-length.
Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

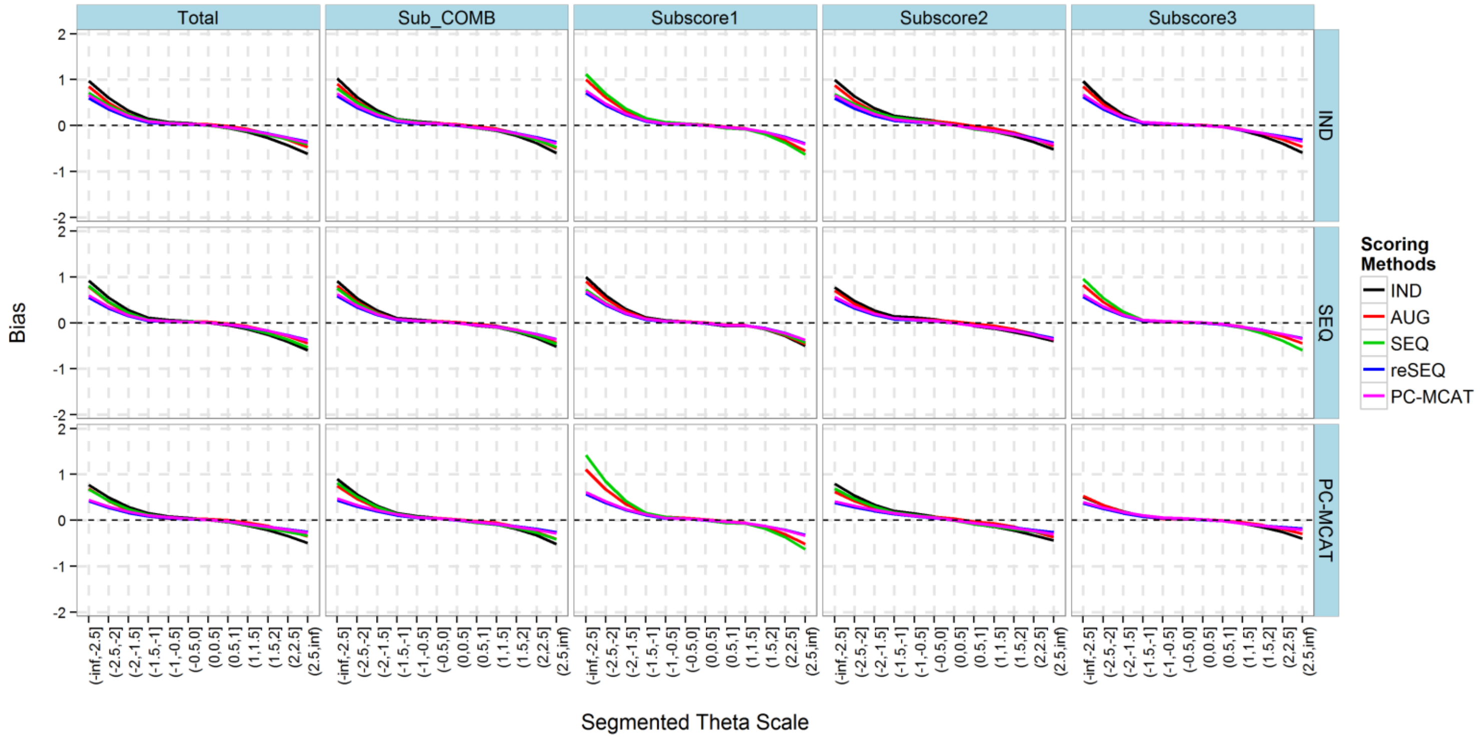

Figure 15. Conditional Bias of $\hat{\theta}$ for All the Conditions with A High Correlation Structure and A 10-Item Sub-length.

Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

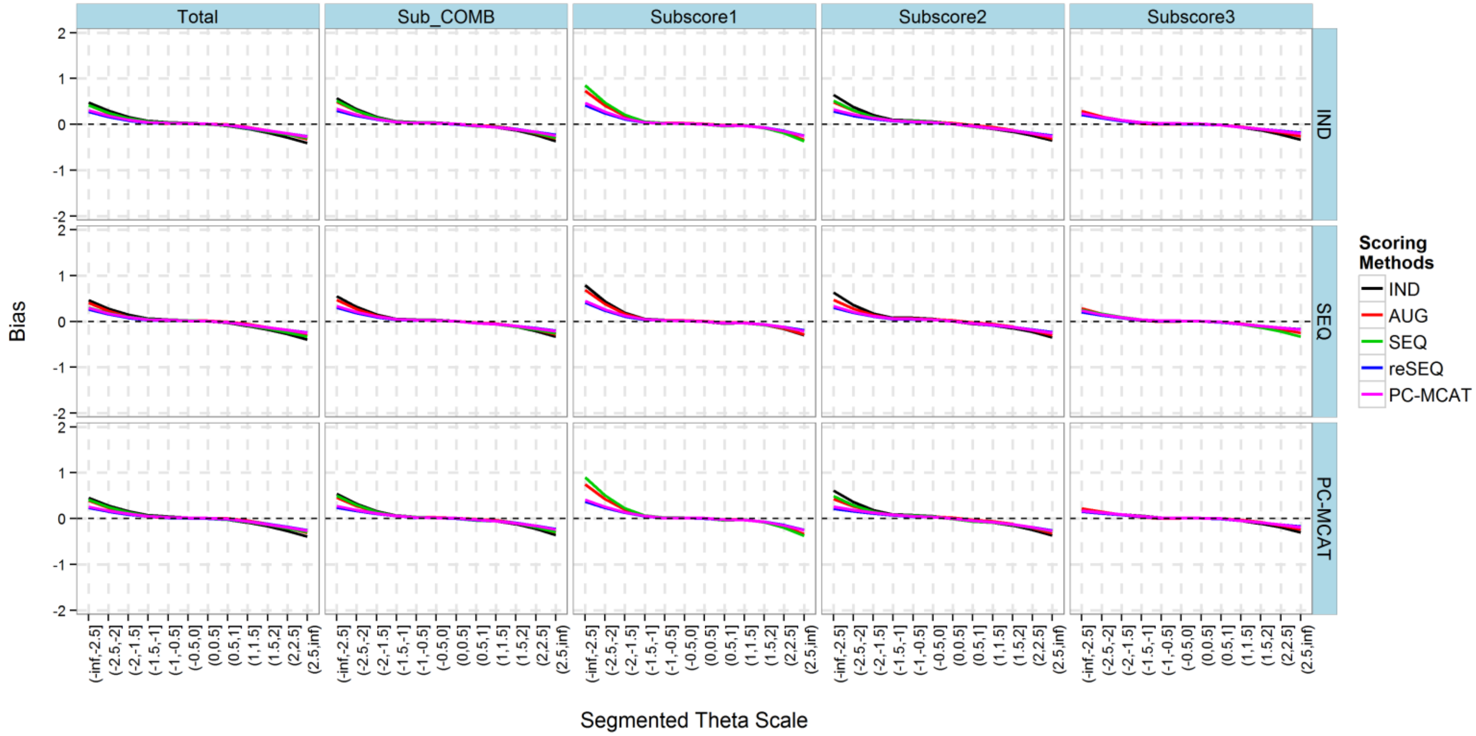

Figure 16. Conditional Bias of $\hat{\theta}$ for All the Conditions with A High Correlation Structure and A 20-Item Sub-length.
Note: The five columns represent the score types; The three rows represent the three item selection algorithms; The five lines in each cell represent the five scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the conditional bias; The points on the x-axis of each cell represent the segmented theta scale intervals; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

APPENDIX E

OUTCOME MEASURES BETWEEN PC-MCAT AND MCAT

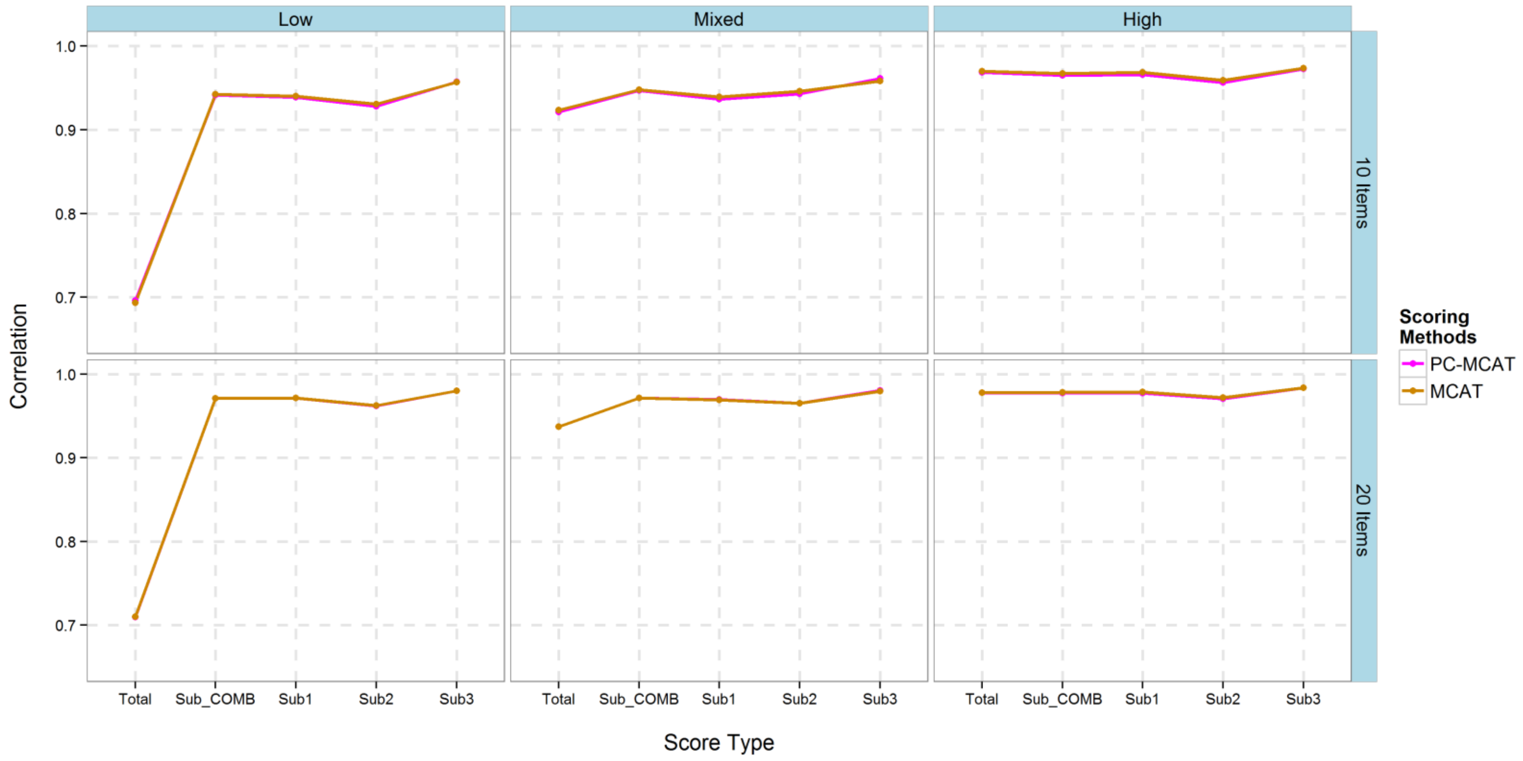

Figure 17. Correlation between θ and $\hat{\theta}$ Yielded by PC-MCAT and MCAT for All the Conditions.
Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The two lines in each cell represent the PC-MCAT and MCAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

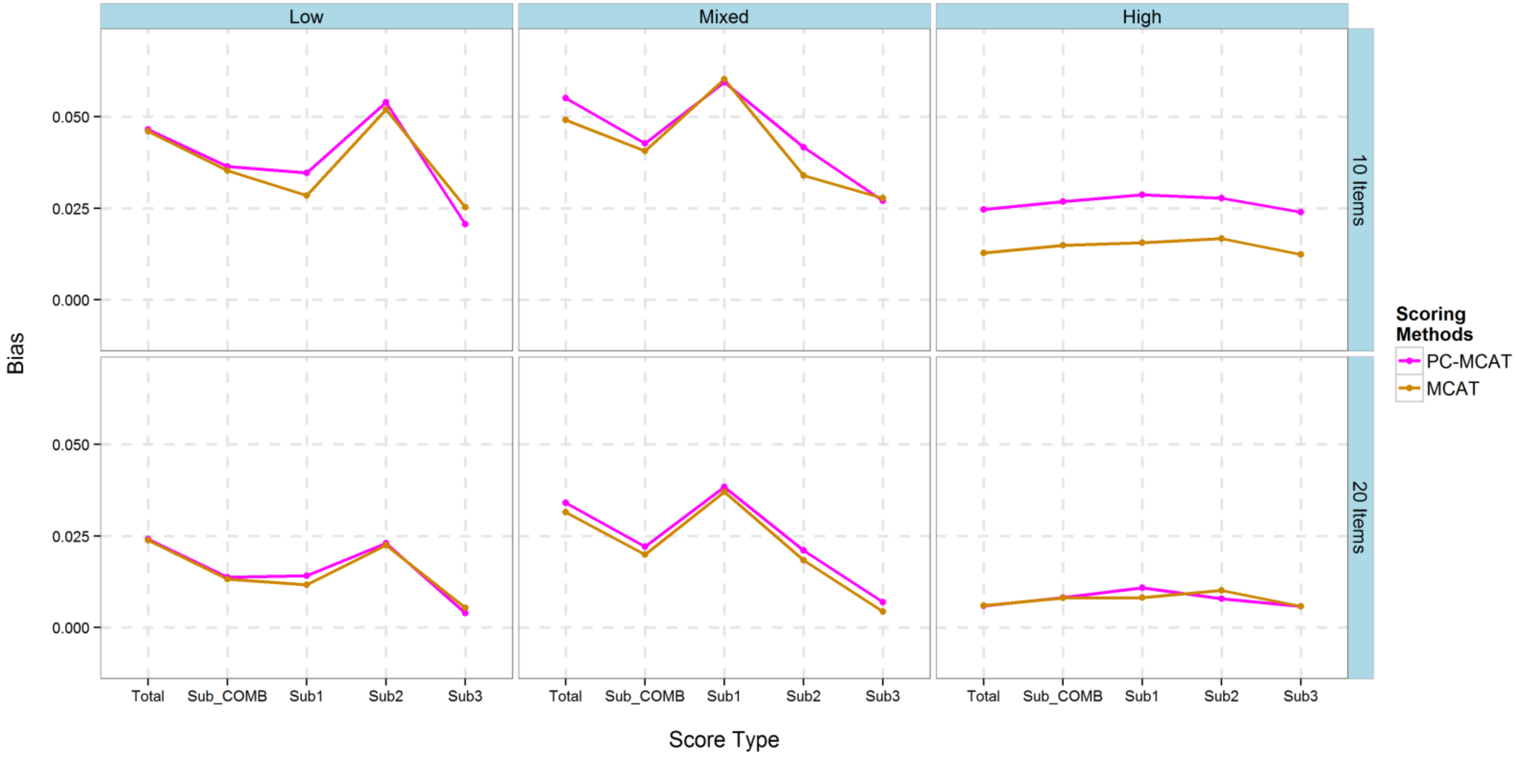

Figure 18. Bias of $\hat{\theta}$ Yielded by PC-MCAT and MCAT for All the Conditions.

Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The two lines in each cell represent the PC-MCAT and MCAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

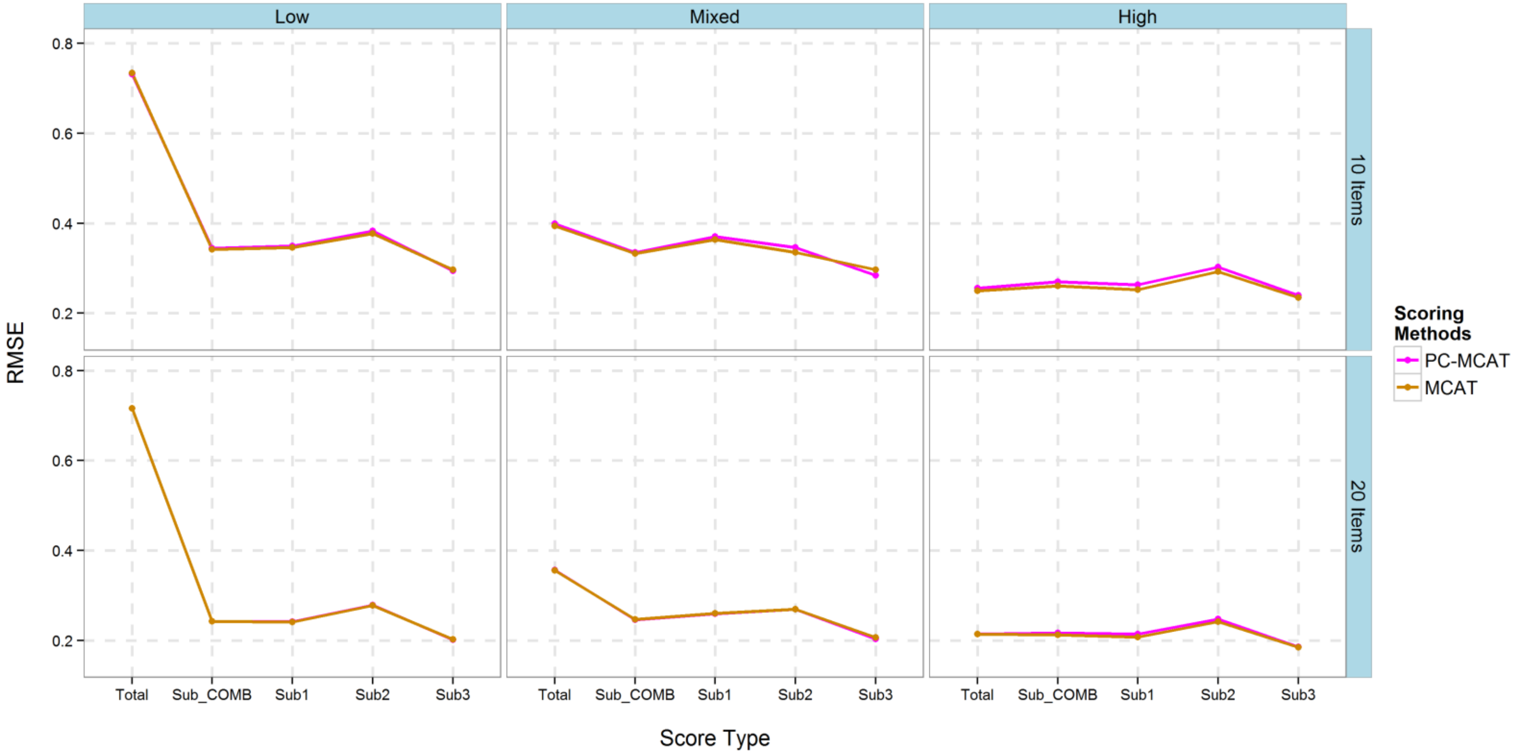

Figure 19. RMSE of $\hat{\theta}$ Yielded by PC-MCAT and MCAT for All the Conditions.
Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The two lines in each cell represent the PC-MCAT and MCAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

Table 26

Percent on the Similarity of Items Selected by PC-MCAT and MCAT

	Low	Mixed	High
10 items	.917	.852	.801
20 items	.956	.922	.893

APPENDIX F

OUTCOME MEASURES AMONG THE FOUR ORIGINAL CAT SCORING METHODS

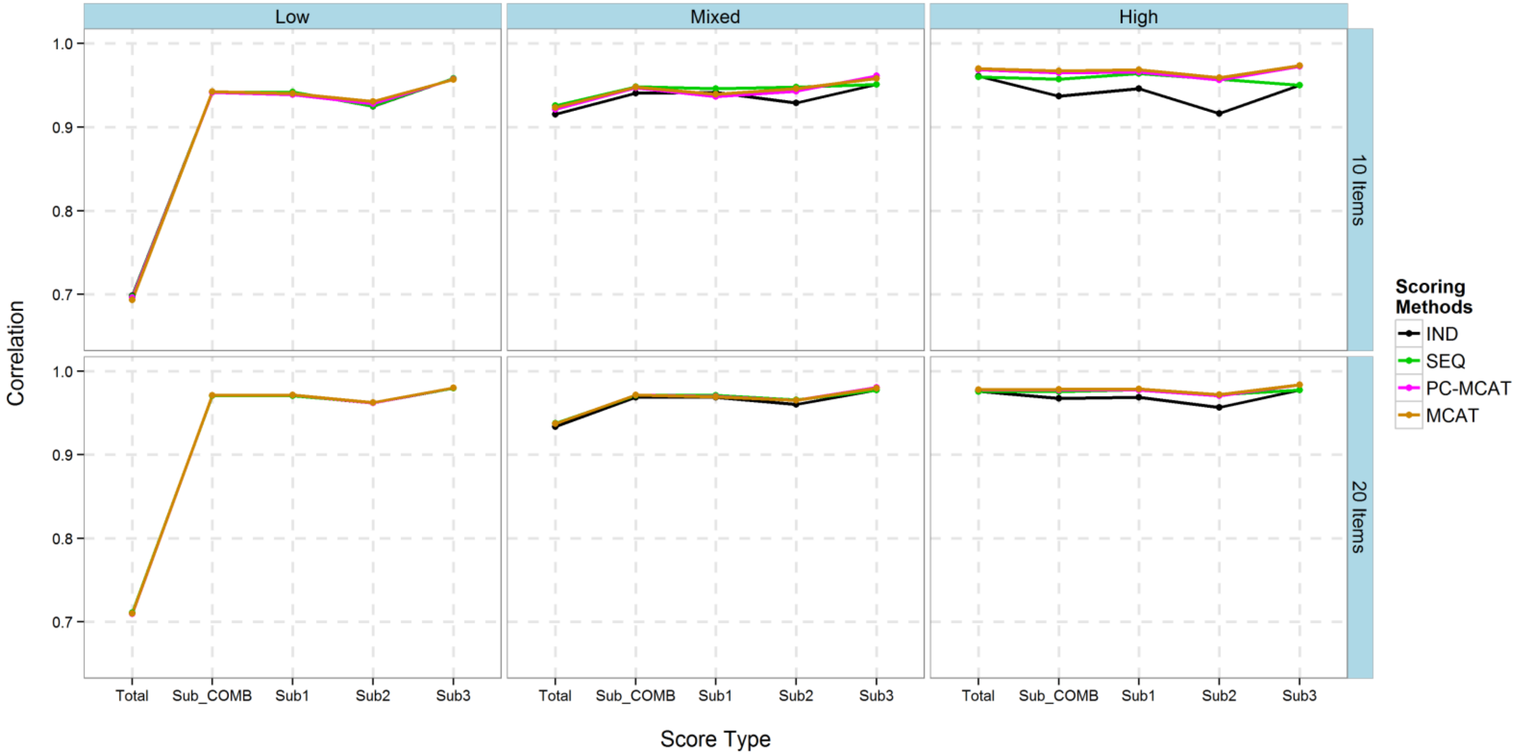

Figure 20. Correlation between θ and $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All the Conditions.

Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The four lines in each cell represent the four original CAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

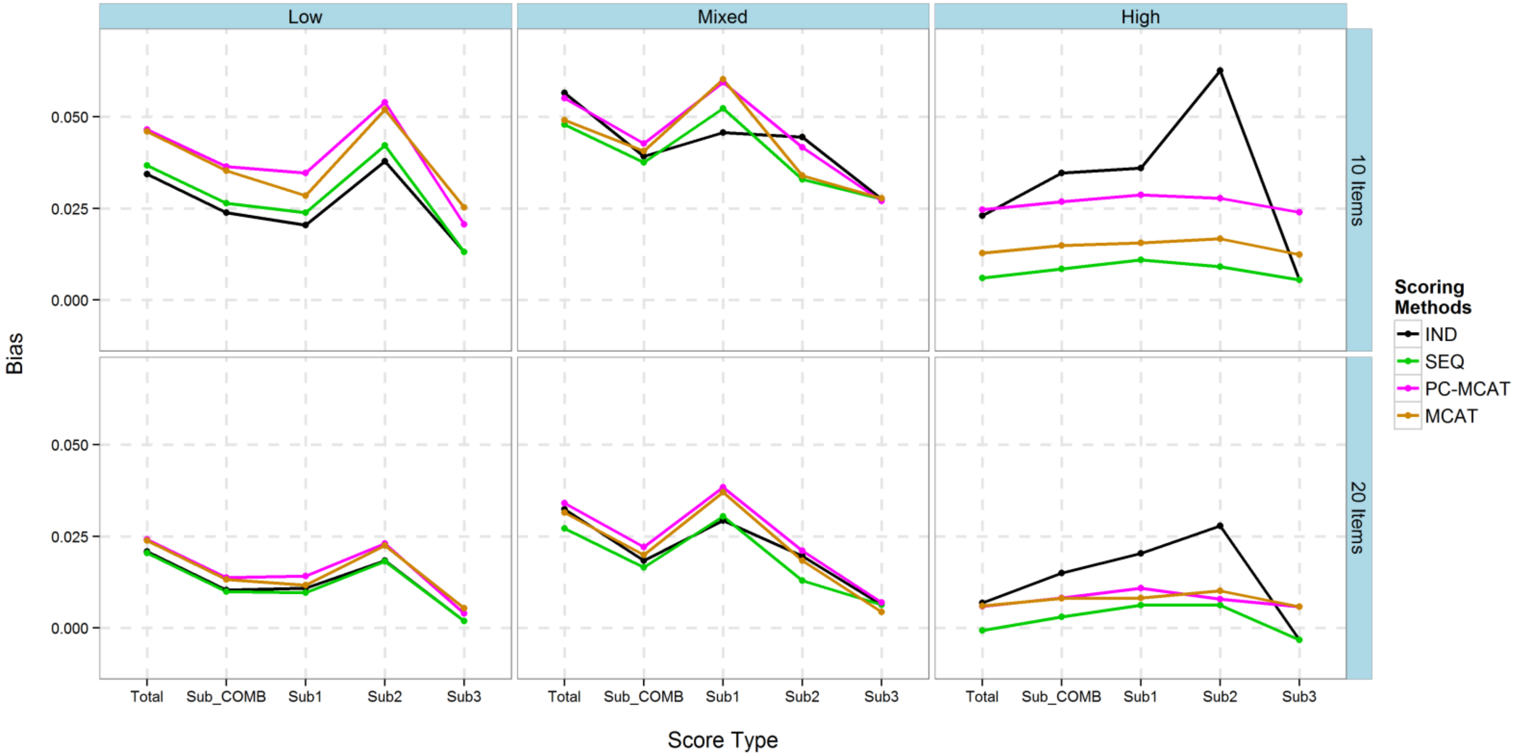

Figure 21. Bias of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All the Conditions.
Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The four lines in each cell represent the four original CAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

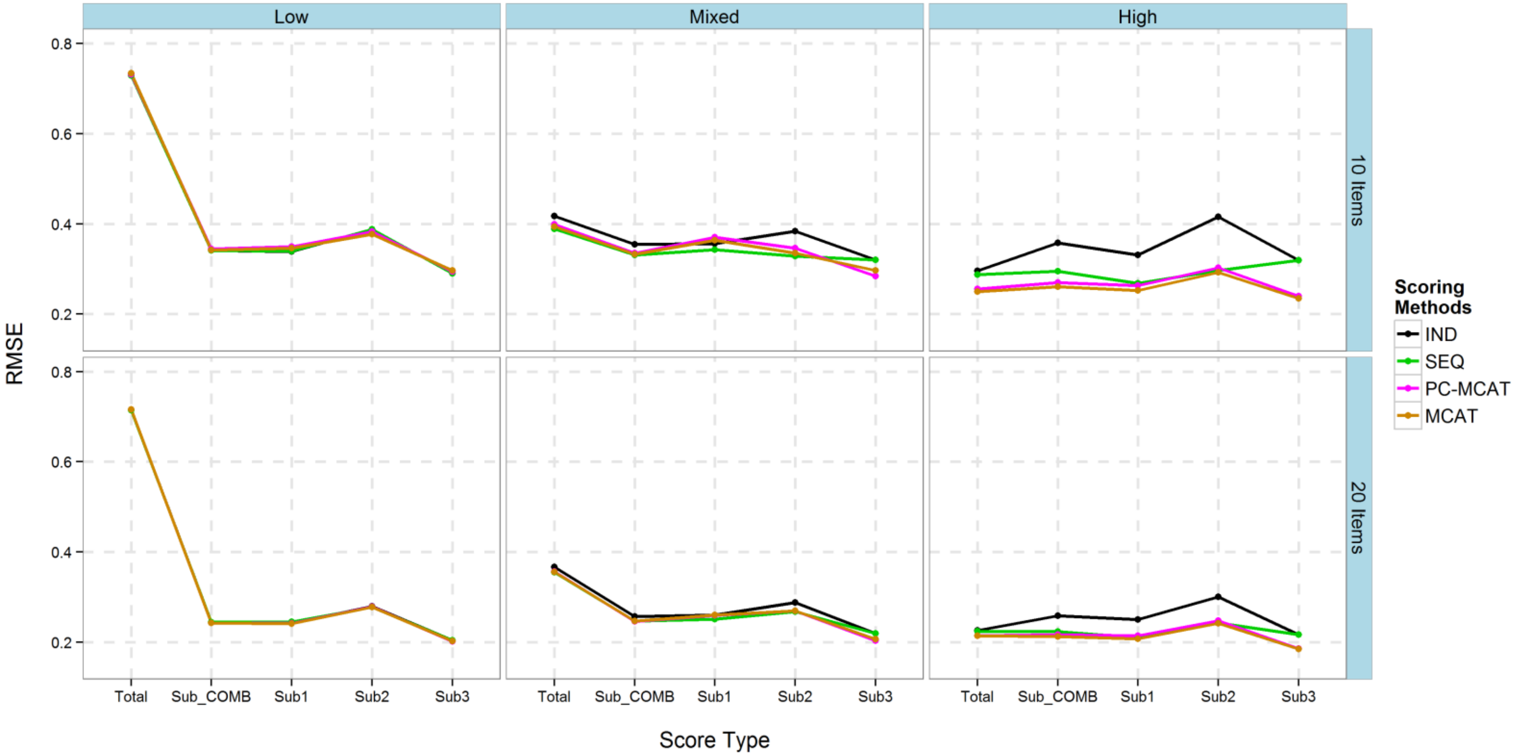

Figure 22. RMSE of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All the Conditions.

Note. The three columns represent the three correlation structures; The two rows represent the two sublengths; The four lines in each cell represent the four original CAT scoring methods, which may be overlapped if the values are all equal or too close; The y-axis in each cell represents the scale of the correlation; The five points on the x-axis of each cell represent the five score types; Sub_COMB refers to the combination of all the three subtests as one test for calculation.

Table 27

Correlation (Difference Values) between θ and $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_	Subtest	Subtest	Subtest	Total	Sub_	Subtest	Subtest	Subtest
			COMB	1	2	3		COMB	1	2	3
Low	IND-CAT	.698	.942	.942	.925	.958	.711	.971	.971	.962	.979
	SEQ-CAT	-.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	PC-MCAT	-.002	.000	-.003	.003	-.001	-.001	.000	.000	.000	.000
	MCAT	-.005	.001	-.002	.006	-.002	-.001	.000	.000	.000	.000
Mixed	IND-CAT	.915	.940	.941	.929	.951	.933	.969	.969	.960	.977
	SEQ-CAT	.010	.008	.005	.019	.000	.004	.002	.002	.005	.000
	PC-MCAT	.006	.006	-.005	.014	.011	.004	.003	.000	.005	.003
	MCAT	.008	.007	-.002	.017	.007	.004	.002	.000	.005	.002
High	IND-CAT	.961	.937	.946	.916	.950	.976	.967	.969	.957	.977
	SEQ-CAT	-.001	.020	.018	.041	.000	.000	.008	.009	.015	.000
	PC-MCAT	.008	.028	.020	.040	.022	.002	.010	.008	.014	.006
	MCAT	.009	.030	.023	.043	.023	.002	.011	.010	.015	.006

Note. The highlighted values are the original values of the correlation between θ and $\hat{\theta}$ estimated by the original IND-UCAT. The other values are the differences between the other three original CAT scoring methods and the original IND-UCAT on correlation; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 30, they were caused by rounding errors.

Table 28

Bias (Difference Values) of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3	Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3
Low	IND-CAT	.034	.024	.020	.038	.013	.021	.010	.011	.018	.002
	SEQ-CAT	.002	.003	.003	.004	.000	.000	-.001	-.001	.000	.000
	PC-MCAT	.012	.013	.014	.016	.008	.003	.003	.003	.005	.002
	MCAT	.012	.011	.008	.014	.012	.003	.003	.001	.004	.003
Mixed	IND-CAT	.057	.039	.046	.044	.027	.032	.018	.029	.020	.006
	SEQ-CAT	-.009	-.002	.007	-.011	.000	-.005	-.002	.001	-.007	.000
	PC-MCAT	-.001	.004	.014	-.003	.000	.002	.004	.009	.001	.001
	MCAT	-.007	.001	.015	-.010	.000	-.001	.002	.008	-.001	-.002
High	IND-CAT	.023	.035	.036	.063	.005	.007	.015	.020	.028	-.003
	SEQ-CAT	-.017	-.026	-.025	-.053	.000	-.006	-.012	-.014	-.022	.000
	PC-MCAT	.002	-.008	-.007	-.035	.018	-.001	-.007	-.010	-.020	.002
	MCAT	-.010	-.020	-.020	-.046	.007	-.001	-.007	-.012	-.018	.003

Note. The highlighted values are the original values of the bias of $\hat{\theta}$ estimated by the original IND-UCAT; The other values are the differences between the other three CAT scoring methods and the original IND-UCAT on bias; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 31, they were caused by rounding errors.

Table 29

RMSE (Difference Values) of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3	Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3
Low	IND-CAT	.729	.341	.339	.388	.290	.715	.244	.243	.279	.204
	SEQ-CAT	.000	.000	.001	-.001	.000	-.001	.000	.002	-.002	.000
	PC-MCAT	.002	.003	.011	-.005	.004	.001	-.001	-.001	-.001	-.002
	MCAT	.004	.000	.007	-.011	.007	.001	-.002	-.002	-.001	-.002
Mixed	IND-CAT	.417	.354	.356	.384	.320	.366	.257	.260	.288	.219
	SEQ-CAT	-.028	-.024	-.014	-.055	.000	-.012	-.010	-.009	-.02	.000
	PC-MCAT	-.018	-.019	.015	-.038	-.036	-.010	-.011	.000	-.019	-.015
	MCAT	-.023	-.021	.008	-.049	-.024	-.010	-.010	.000	-.019	-.012
High	IND-CAT	.295	.358	.331	.416	.319	.226	.258	.250	.300	.217
	SEQ-CAT	-.008	-.063	-.063	-.119	.000	-.002	-.035	-.040	-.058	.000
	PC-MCAT	-.040	-.088	-.067	-.113	-.079	-.012	-.041	-.036	-.053	-.031
	MCAT	-.046	-.097	-.079	-.123	-.084	-.012	-.046	-.043	-.059	-.032

Note. The highlighted values are the original values of the RMSE of $\hat{\theta}$ estimated by the original IND-UCAT; The other values are the differences between the other three CAT scoring methods and the original IND-UCAT on RMSE; Positive difference values mean higher than the highlighted values and negative difference values mean lower than the highlighted values; Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 32, they were caused by rounding errors.

Table 30

Correlation (Original Values) between θ and $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3	Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3
Low	IND-CAT	.698	.942	.942	.925	.958	.711	.971	.971	.962	.979
	SEQ-CAT	.697	.942	.942	.925	.958	.711	.971	.970	.962	.979
	PC-MCAT	.696	.941	.939	.928	.957	.710	.971	.971	.962	.980
	MCAT	.693	.942	.940	.930	.957	.710	.971	.971	.962	.980
Mixed	IND-CAT	.915	.940	.941	.929	.951	.933	.969	.969	.96	.977
	SEQ-CAT	.925	.948	.946	.948	.951	.938	.971	.971	.965	.977
	PC-MCAT	.921	.947	.937	.943	.962	.937	.971	.969	.965	.980
	MCAT	.923	.948	.939	.946	.958	.937	.971	.969	.965	.980
High	IND-CAT	.961	.937	.946	.916	.950	.976	.967	.969	.957	.977
	SEQ-CAT	.960	.957	.964	.957	.950	.976	.976	.978	.972	.977
	PC-MCAT	.968	.965	.966	.956	.973	.978	.977	.977	.971	.983
	MCAT	.970	.967	.969	.959	.974	.978	.978	.978	.972	.984

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 27, they were caused by rounding errors.

Table 31

Bias (Original Values) of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3	Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3
Low	IND-CAT	.034	.024	.020	.038	.013	.021	.010	.011	.018	.002
	SEQ-CAT	.037	.026	.024	.042	.013	.020	.010	.010	.018	.002
	PC-MCAT	.047	.036	.035	.054	.021	.024	.014	.014	.023	.004
	MCAT	.046	.035	.028	.052	.025	.024	.013	.012	.022	.005
Mixed	IND-CAT	.057	.039	.046	.044	.027	.032	.018	.029	.020	.006
	SEQ-CAT	.048	.038	.052	.033	.027	.027	.017	.030	.013	.006
	PC-MCAT	.055	.043	.059	.042	.027	.034	.022	.038	.021	.007
	MCAT	.049	.041	.060	.034	.028	.031	.020	.037	.018	.004
High	IND-CAT	.023	.035	.036	.063	.005	.007	.015	.020	.028	-.003
	SEQ-CAT	.006	.008	.011	.009	.005	-.001	.003	.006	.006	-.003
	PC-MCAT	.025	.027	.029	.028	.024	.006	.008	.011	.008	.006
	MCAT	.013	.015	.016	.017	.012	.006	.008	.008	.010	.006

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 28, they were caused by rounding errors.

Table 32

RMSE (Original Values) of $\hat{\theta}$ Yielded by the Four Original CAT Scoring Methods for All Conditions

		10-Item Sublength					20-Item Sublength				
		Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3	Total	Sub_ COMB	Subtest 1	Subtest 2	Subtest 3
Low	IND-CAT	.729	.341	.339	.388	.290	.715	.244	.243	.279	.204
	SEQ-CAT	.729	.341	.340	.388	.290	.714	.244	.244	.277	.204
	PC-MCAT	.731	.344	.349	.383	.294	.716	.243	.242	.278	.202
	MCAT	.734	.341	.346	.377	.297	.716	.242	.241	.278	.202
Mixed	IND-CAT	.417	.354	.356	.384	.320	.366	.257	.26	.288	.219
	SEQ-CAT	.388	.331	.342	.329	.320	.354	.247	.251	.268	.219
	PC-MCAT	.399	.335	.371	.346	.284	.356	.245	.259	.269	.203
	MCAT	.394	.333	.364	.335	.296	.356	.247	.260	.269	.207
High	IND-CAT	.295	.358	.331	.416	.319	.226	.258	.250	.300	.217
	SEQ-CAT	.287	.295	.268	.297	.319	.224	.223	.210	.242	.217
	PC-MCAT	.256	.270	.263	.302	.240	.214	.217	.214	.247	.185
	MCAT	.249	.261	.252	.292	.235	.214	.212	.207	.242	.185

Note. Sub_COMB refers to the combination of all the three subtests as one test for calculation; If some discrepancies occurred between this table and Table 29, they were caused by rounding errors.