DIAGNOSTIC TOOLS FOR IMPROVING THE AMOUNT OF ADAPTATION IN ADAPTIVE TESTS USING OVERALL AND CONDITIONAL INDICES OF ADAPTATION

By

Unhee Ju

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2019

ABSTRACT

DIAGNOSTIC TOOLS FOR IMPROVING THE AMOUNT OF ADAPTATION IN ADAPTIVE TESTS USING OVERALL AND CONDITIONAL INDICES OF ADAPTATION

By

Unhee Ju

In recent years, computerized adaptive testing (CAT) has been widely used in educational and clinical settings. The basic idea of CAT is relatively straightforward. A computer is used to administer items tailored for individuals to maximize the measurement precision of their proficiency estimates. However, the administration of CAT is not so simple. Those who administer CATs must, while trying to optimize an item selection criterion, consider a variety of practical issues such as test security, content balancing, the purpose of testing, and other test specifications. Such extraneous factors make it possible that a CAT might have so many constraints that in practice it is barely adaptive at all. This concern is at the forefront of the current study, which poses two key questions: How adaptive is a highly adaptive test really? How can the level of adaptation be improved?

This study aims to develop three new statistical indicators to measure the amount of adaptation conditional on the examinees' proficiency levels in CAT. It also aims to evaluate the feasibility and utility of these adaptation measures in helping to diagnose and improve adaptivity that occurs during the CAT administration. Extending work done by Reckase, Ju, and Kim (2018), the proposed measures are based on three components—the differences in the locations between the selected items and the examinee's current proficiency estimates, the variations in the item locations administered to each examinee, and the magnitude of information that the test presents to each examinee. Hence, they can be used to assess adaptivity during the CAT process,

as well as to identify differences in the level of adaptation for individuals or subgroups of examinees.

To demonstrate the performance of the proposed adaptation indices, this study conducted analyses of real operational testing data from a healthcare licensure examination, as well as comprehensive simulation studies under various conditions that affect adaptivity in a CAT. The key findings of the study suggest that the proposed adaptation indices are likely to function as intended to sensitively detect the magnitude of adaptivity for a CAT over the proficiency continuum. These new measures shed light on how much adaptation of a given test occurs across individual proficiency levels or subpopulations. With some guidelines for the interpretation of these measures recommended in this study, the adaptation indices can also readily serve as diagnostic tools in practice for helping test practitioners design item pools and adaptive tests that support high adaptivity.

ACKNOWLEDGEMENTS

I would like to express deep gratitude to my advisor and committee chair, Dr. Mark D. Reckase. The support, guidance, and encouragement that he has provided throughout my doctorate training years has been priceless. He has given me numerous opportunities to conduct research with him, showing me I could enjoy doing research and with self-motivation. A passionate scholar and wise educator, he has been a great role model to me. I would not have come this far without his tremendous academic support and emotional encouragement.

For their support and invaluable comments, I also sincerely appreciate my committee members, Dr. Kimberly Kelly, Dr. Richard Houang, and Dr. Christopher Nye. I thank National Council of State Boards of Nursing (NCSBN), especially Dr. Qian Hong for allowing me to have access to operational data used for this dissertation study. Also, I am deeply thankful to Dr. Carl F. Falk for sharing his knowledge and research experience, as well as to Dr. Eunsoo Cho for providing financial support the last two years and research opportunities in applied research areas. I also want to thank my advisor in South Korea, Dr. Eunlim Chi who first sparked my interest in Educational Measurement (Psychometrics) and took me under her wing until the end of my PhD journey.

My special thanks go out to Nancy Duchesneau and William Sullivan for reading over my dissertation, and to my close colleagues and friends at Michigan State University who have helped me stay steady and throughout the six years always stood by me—especially Jiahui Zhang (I’m really lucky to have had you from the beginning to the end of this journey!), Ina Choi, Jihyun Park, Ajin Lee, and Susie Kim. I’m so grateful to all of you for showing me true friendship, cheering me on, and being with me through every important stage of this journey.

Most importantly, I dedicate this dissertation to my family. No words can fully express my heartfelt gratitude and appreciation to my mom and dad, Taesook Kang and Yeonghwan Ju, and my brother, Bongseop for their unconditional love, patience, confidence, and belief in me. None of this would have been possible without your love, support, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1.

INTRODUCTION

1.1 Background

Computerized adaptive testing (CAT) has been used in a wide range of settings. These include licensure and certification examination (e.g., National Council Licensure Examination [NCSBN, 2016]), admissions tests (e.g., Graduate Record Examinations®, Graduate Management Admission Test), achievement assessments within statewide educational system (e.g., Minnesota [Minnesota Department of Education, 2017]), and clinical settings to assess psychological or health-related outcomes (e.g., anxiety- and depression-CAT [Walter, 2010]), and still others. The popularity of CAT is attributed to its merits of efficient testing and high measurement precision of proficiency estimates. As CAT is implemented, it selects, administers, and scores items tailored for each individual, based on optimizing criteria such as maximizing the Fisher information at the current proficiency estimate.

The basic idea of CAT is relatively straightforward. However, numerous practical challenges to the deployment of CAT have persisted. These concern the design, implementation, and maintenance of a CAT program with respect to development and maintenance of the item pool, test administration (e.g., item selection, scoring, and termination procedures), test security, and examinee issues. The success of a CAT program is dependent on how well these practical concerns are addressed (see Wise & Kingsbury, 2000, for details). Measurement professionals have resolved a number of these issues in CAT using alternative options. For instance, a variety of appropriate constraints are imposed on item selection to conform to test specifications (e.g., Kingsbury & Zara, 1989; van der Linden & Reese, 1998) and item exposure control

requirements (e.g., Chang, Qian, & Ying, 2001; Sympson & Hetter, 1985). Some CATs adapt at the testlet level to incorporate the grouped items associated with a common stimulus (e.g., Wainer & Kiely, 1987). Multistage testing (MST), a special version of CAT, adapts at the stage level using pre-constructed modules, allowing reviews on psychometric and content properties and more efficient handling of complex test constraints (e.g., Yan, von Davier, & Lewis, 2016). In addition, CATs differ in their item pool design, their stopping rule, and estimation procedures. All of these features have influence on an operational CAT program.

Although many of these designs and variations in the implementation of CATs are given the label "adaptive tests," questions arise about whether they would be equally adaptive to examinees' proficiency. An administered CAT may not be very adaptive if it imposes too many constraints on item selection for the purpose of strong exposure control and strict content-balancing using a small item pool with limited spread in item difficulty. A severely constrained CAT may lead to all examinees getting almost the same test, making it nearly the same as paper-and-pencil tests. If a CAT has a relatively large item pool, however, without any constraints on item selection, examinees may receive an optimal set of items customized for their proficiency levels during a test, showing a high level of adaptation. Another issue for the consideration is the test fairness for examinees. Sometimes, a testing program uses multiple item pools for the test administration. In this situation, some examinees receive the items given in the test that match well with the students' proficiency levels from the pool of high-quality items, while others could take the items less adapted for their abilities due to the different item pool characteristics assembled from the master item pool.

Spurred by such concerns, Reckase and colleagues (Reckase, Ju, & Kim, 2018) proposed three adaptation indices to quantify the amount of adaptation that occurs in CAT based on the

variance of the difficulty parameters for the items administered to the examinees. While these measures are useful and work fairly well (e.g., Reckase, Ju, & Kim, 2017, 2019), they are limited to the evaluation of the adaptivity of tests over the entire group of examinees, rather than for individuals or subgroups. The measures give us the overall diagnosis of the adaptivity of administered tests but provide no specific information about the degree of the adaptivity conditional on the examinee' proficiency level. The latter information would be useful to modify test designs or the quality of the item pool for reaching the optimal adaptation desired for the test's purpose. In addition, the overall indices introduced by Reckase et al. (2018) focus on how appropriately the items administered to examinees are customized to their *final* proficiency estimates, while item selection is driven by *interim* proficiency estimates. The adaptation indices thus do not know the quality of adaptation during the intermediate stages of CAT because the final proficiency estimate is not known until the end of the test administration. In other words, these adaptation measures conduct the post-evaluation of whether the items presented in the test are optimal for the examinee's final proficiency levels, but they are blind to whether the test provides the items that well match their momentary proficiency estimates during the CAT process.

In response to this perceived necessity, Kingsbury and Wise (2018) suggested a new measure of adaptation based on item response theory (IRT) test information. Although this index is informative, it fails to take into account the alignment of and variations in difficulty for the items administered to each examinee, as well as it focuses on the actual test information based on the final proficiency estimates. Plus, as found in Reckase et al. (2018, 2019), a single index may be insufficient to assess all the relevant information about the magnitude of adaptation that

happens during a CAT because adaptivity is intertwined with item pools, item selection algorithms, proficiency estimators, and other test specifications.

1.2 Research Questions

To address practical needs and the gap in CAT literature, this study proposes new statistical indicators to examine the level of adaptation conditional on an examinee's proficiency using (1) the locations of the items (item difficulty or the location of the maximum information) administered to each examinee, (2) their variances, and (3) their IRT information. The study then explores the capabilities of these three new adaptation measures as tools for understanding how well adaptivity occurs in the applications of CATs, as well as demonstrates the practical utility of these indices using real operational data from the licensure and certification examination. Consequently, a new class of the adaptivity indices introduced here can help measurement professionals and test developers understand the adaptation costs associated with item pool designs, test designs, constraints on item selection, and so forth. Overall, this study is guided by five research questions:

1. How sensitive are the conditional adaptation indices to changing characteristics of the item pool, proficiency estimators, and IRT models?
2. For a given population of examinees, test specifications, and an item pool, how can the conditional adaptation measures be used to revise the item pool in such a way that a CAT works better?
3. Do the conditional adaptation indices capture the varying degree of adaptivity resulting from constraints imposed on item selection for exposure control?
4. Can the conditional adaptation indices be used to gauge the amount of adaptation incurred by adaptive test designs (fully adaptive tests vs. multistage adaptive tests)?

5. Do the conditional adaptation indices function appropriately to diagnose the adaptivity of an operational CAT program?

The first four research questions are answered through comprehensive simulation studies, and the last research question is demonstrated using operational variable-length CAT data. The next chapter reviews the features of the item response theory (IRT) models, the components of CAT, and factors that possibly affect the amount of adaptation. Chapter 3 describes indices to measure the amount of adaptation, followed by a chapter that gives details about simulation designs and real operational data of the CAT program. Finally, the last two chapters (Chapter 4 and Chapter 5) present the findings for the performance of the proposed indices and discuss how the new set of adaptation measures can be efficiently and directly utilized for comparing the quality of adaptivity at individuals or subpopulations and for improving the amount of adaptation by revising the item pools or the test designs and specifications.

CHAPTER 2.

LITERATURE REVIEW

This literature review chapter consists of three main sections. The first section explores the characteristics of item response theory (IRT), which is the fundamental basis of computerized adaptive testing (CAT) in terms of scoring and item selection. The second section summarizes the components of CAT. The last section discusses plausible factors that have influences on the amount of adaptation for CAT.

2.1 Item Response Theory

Item response theory (IRT; Lord, 1980) describes the interaction between test items and examinees through a mathematical model, called an *item response function* (IRF) that specifies the probability of a correct response on a given item, with item parameters, as a function of an examinee's proficiency (θ). Item parameters, in general, include (1) an item difficulty parameter that indicates the relative difficulty or easiness of an item (i.e., location parameter) to examinees, (2) an item discrimination parameter that describes how well an item distinguishes between examinees of varying proficiency levels, and (3) an item pseudo-guessing parameter that indicates the possibility of giving a correct answer by chance.

IRT models are usually classified into two types based on how the item responses are scored: dichotomous IRT models and polytomous IRT models. Since the study focuses on the tests of dichotomously scored items, this section only describes dichotomous IRT models, which are commonly applied to binary scored multiple-choice (MC) items or true/false items (e.g., correct/incorrect). Three frequently-noted dichotomous IRT models include the one-parameter

logistic (1PL) or Rasch model (Rasch, 1961), the two-parameter logistic (2PL) model (Birnbaum, 1968), and the three-parameter logistic (3PL) model (Birnbaum, 1968).

2.1.1 Rasch (1PL) model

The 1PL model, also known as the Rasch model, is the most parsimonious model among the commonly considered IRT models. It assumes unit discrimination for all items and no guessing. The IRF of the Rasch model specifies the probability of a correct response on item *i* for examinee *j* by:

$$P_{ij}(u_{ij} = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \tag{2.1}$$

where θ_j represents the proficiency level of examinee *j*, and b_i denotes the item difficulty parameter of item *i*.

2.1.2 2PL model

The 2PL model considers not only the item difficulty (b_i) but also the item discrimination parameter (a_i). However, it does not assume that guessing contributes to the examinee's response on an item. In this model, the probability of a correct response on item *i* administered to examinee *j* can be defined as:

$$P_{ij}(u_{ij} = 1|\theta_j) = \frac{\exp\left(a_i(\theta_j - b_i)\right)}{1 + \exp\left(a_i(\theta_j - b_i)\right)}. \tag{2.2}$$

2.1.3 3PL model

Unlike the 2PL model, the 3PL model considers that, for an examinee with very low proficiency, there is a possibility that the examinee correctly answers an item through a pseudo-guessing parameter (c_i), especially by chance with a MC item. The probability of examinee *j* having a correct response for item *i* is:

$$P\left(u_{ij}=1\middle|\theta_j\right)=c_i+(1-c_i)\frac{\exp\left(a_i\left(\theta_j-b_i\right)\right)}{1+\exp\left(a_i\left(\theta_j-b_i\right)\right)}. \quad (2.3)$$

2.1.4 Information function for dichotomous IRT models

IRT provides a measurement precision for the items over the proficiency continuum through the usage of an item information function. The IRT information function of dichotomously scored items can be expressed as follows (Lord, 1980):

$$I(\theta)=\frac{P'(\theta)^2}{P(\theta)(1-P(\theta))}, \quad (2.4)$$

where $P(\theta)$ is the probability of correctly answering the item given θ, and $P'(\theta)$ is the first derivative of the probability function.

For the 3PL logistic model, Equation 2.4 can be represented as:

$$I(\theta)=a^2\frac{Q(\theta)}{P(\theta)}\left[\frac{P(\theta)-c}{1-c}\right]^2, \quad (2.5)$$

where $Q(\theta)=1-P(\theta)$. Based on Equation 2.5, the information function for the 2PL model can be obtained by setting $c=0$, and the information function for the Rasch model is obtained by setting $c=0$ and $a=1$. High information for an item at a particular proficiency level indicates that the item is very informative to precisely measure the examinee's proficiency. The amount of information is greatly affected by *a*-parameter. Note that for the 1PL model, the items have the same amount of maximum information at the location where the proficiency is close to the *b*-parameter of the item (i.e., 0.25). In addition, the test information function can be simply calculated by summing the information functions for the test items contingent on proficiency.

2.2 Computerized Adaptive Testing

In recent years, with discussion about visions for next-generation assessment, CAT has received great re-attention in educational system for personalized assessments (e.g., Conely, 2018; Embretson, 2001). CAT delivers an individualized test tailored to a test-taker, and thus it can shorten the test length without sacrificing measurement precision. Compared to paper and pencil (P&P) linear tests, the advantages of CAT reported in the literature (e.g., Chang, 2004; Gibbons et al., 2008; Meijer & Nering, 1999; van der Linden, 2010) include shorter tests, improved test reliability, and immediate test scoring and reporting. Also, CAT allows one to obtain information that are not available in P&P tests, including response time (e.g., Wise, Bhola, & Yang, 2006), graphical entries, mouse/eye movements, and so forth, which may open new avenues for future research that helps understand examinee's testing behaviors or cognitive activities. Furthermore, CAT enables the use of a variety of innovative items and technology-enhanced items, which leads to improvement in the validity evidence of tests that cannot be obtained in P&P tests (Luecht & Clauser, 2002).

Basically, the CAT algorithm starts with a selection of the first item whose *b*-parameter is matched with a pre-determined initial proficiency estimate. After the item is scored and the examinee's proficiency estimate is updated, the next item is then selected at the current proficiency estimate from the given item pool based on the item selection criterion. This procedure continues until a stopping rule is satisfied. Reckase (1989) reported four core components for an operational CAT: item pool, item selection procedure, scoring procedure, and stopping rules. Constraints for content balancing and exposure control are considered in the item selection procedure. In what follows, these four components will be briefly illustrated.

2.2.1 Item pool

A paramount element that affects the performance of a CAT in numerous ways is the item pool. For instance, the item pool affects the proficiency estimates, eventually influencing a subsequent item to be administered. In real operational settings, there are two types of item pools for CAT. One is a *master pool*, called "vat", which consists of the large number of items to supply the testing program. The other is an *operational item pool*, which is used during a testing implementation period to provide items tailored to the individuals' proficiency levels. A testing company typically assembles the operational item pools from the master pool so as to renew the item pools after a certain period of time usages or a certain number of students take the test using the same item pool.

Without a well-designed item pool, a CAT cannot be successfully implemented. Thus, the size and the quality of the item pool is very essential. The desired item pool for CAT has been recommended to include an adequate number of good quality of items to provide informative tests to the sample of examinees (Flaugher, 2000; McBride, 1977). Here, the good quality of items (i.e., optimal items) generally have high item discriminations (e.g., $a > .08$) and low guessing parameters (e.g., $c < .03$). At the same time, the range of item difficulty in the item pool should cover the distribution of examinee's proficiency levels to ensure that all examinees can take items well-tailored to their proficiency levels (Mills & Stocking, 1996; Urry, 1977). In addition to the statistical requirements of the optimal item pool, the pool should contain items that measure the intended construct for the testing purpose and the use of the test scores (Kane, 2013).

Some research (e.g., Gu & Reckase, 2007; He & Reckase, 2013; Reckase, 2010; Veldkamp & van der Linden, 2010) introduced approaches to design the item pool for CAT but

with different definitions of an optimal item pool. Veldkamp and van der Linden (2010) proposed a method for designing an optimal blueprint for a CAT item pool with the integer programming model that minimizes an estimate of item-writing costs using the classification table defined by item attributes figuring in test specifications (e.g., content, format, word count, item difficulty). The goal of this item pool design process is to figure out the number of items required for each cell of the classification table, guiding the item writing process. However, this method uses the characteristics of a previous or existing item pool as a starting point to define item-writing costs.

Another line of research on item pool design (e.g., Gu & Reckase, 2007; He & Reckase, 2013; Mao, 2014) has been based on the bin-and-union method (Reckase, 2010) with more emphasis on the psychometric properties of an optimal item pool, in lieu of the item-writing costs. It also does not require pre-existing information about the item pool. An optimal item pool defined in this method should include a desired item available for every stage of item selection that matches the current proficiency estimate for each examinee. The optimal item pool is determined by tallying the location of the sequential proficiency estimates for each examinee with the expectation that there would be an item in the pool whose information peaks at that location on the proficiency scale. As the items used for a single examinee can be used for other examinees, the full item pool is determined from the union of the items required for the entire set of examinees of interest (see Reckase (2010) and He and Reckase (2013) for more details of the process). Exposure control, content balancing, and other specifications can be incorporated into the CAT simulations to identify the design for the optimal item pool. I employed this bin-and-union method to design the ideal item pool in Section 4.4.1.1.

The item pool size, another important aspect of the item pool, is dependent on the testing purpose, the CAT specifications (e.g., exposure control and content balancing), and the examinee's proficiency distribution (Parshall, 2002; Reckase, 2010). Prior research (e.g., Chen, Ankenmann, & Spray, 2003; Gönülates, 2015) has generally supported that the size of the item pool should be 10 to 12 times larger than the test length (Stocking, 1994). To investigate the effect of item pool size on adaptivity for CAT, I manipulate the item pool size as a factor in the simulation studies (see Section 4.2.2.1).

2.2.2 Item selection procedure

Another key component of CAT is the item selection algorithm. The most frequently used item selection algorithm is *maximum Fisher information* (MFI; Lord, 1977) due to its easy implementation. The MFI select the next item that has the maximum information in Equation (2.4) at the current proficiency estimate from the available item pool. In other words, it selects the item that most precisely measures the current proficiency estimate. Thus, as this algorithm selects the most informative item, the efficiency of CAT also increases. However, MFI has some disadvantages. At the beginning stages of CAT, there is not enough information of an examinee's proficiency location to guide the MFI item selection procedure, resulting in selected items that might not be the best ones. The items administered earlier in the test could also bring about big jumps in the proficiency estimates. This is the reason why it is recommended for examinees to be extra careful while answering the first few questions of the test. Such issue could be mitigated by using prior information to select items or using different item selection rules such as the Kullback-Leibler measure (Chang & Ying, 1996) at least at the early stages of CAT, especially for a short test (Chen, Ankenmann, & Chang, 2000).

Owen's Bayesian item selection approach (Owen, 1975) is also commonly used in CAT programs. This approach selects the items that minimize the expected posterior variances of the proficiency estimates. To calculate the posterior distribution of the proficiency for item selection in CAT, Owen used a normal approximation with closed-form expressions, instead of the true posterior, in order to minimize the computational complexity. He proved that as the number of the administered items become infinite, the expected value of the posterior distribution will converge to the true value of proficiency. In general, an examinee receives the first item that matches well with the initial proficiency estimate that is equal to the expected value of the prior distribution. The algorithm then searches for a next item that will reduce the posterior variance the most. After each item is administered, a new posterior distribution is computed using the response strings and the prior distribution (usually, normal distribution), and then this updated posterior becomes the prior distribution for selecting the next item. As Owen's Bayesian method is an approximate empirical Bayes procedure for CAT which requires simpler computation, this method is faster than other Bayesian item selection approaches.

In addition to these two item selection approaches, there are other Bayesian item selection procedures. For instance, van der Linden (1998a) proposed several Bayesian item selection criteria based on the full posterior, including maximum posterior-weighted information (MPWI), maximum expected information (MEI), minimum expected posterior variance (MEPV), maximum expected posterior weighted-information (MEPWI). Penfield (2006) compared the performance of MEI and MPWI to MFI, reporting that the Bayesian procedures yielded slightly more precise estimates than MFI. Prior studies (e.g., Choi & Swartz, 2009) also found that these Bayesian item selection procedures are computationally intensive but produce

comparable results to the simpler MFI procedure. Therefore, the MFI procedure is the most widely used in item selection of CAT and used in this dissertation, as well.

Other practical considerations are made in item selection to address the issues of over- or under-exposed items, content validity, students' fairness, and item characteristics for CAT. To handle these practical issues, constraints are generally imposed on the item selection procedures. The constraints on item selection include but are not limited to exposure control, content balancing, and item enemies (Eignor, Stocking, Way, & Steffen, 1993; Weiss, 2011). Among these numerous constraints, the following selections will briefly discuss some constraints on item selection for exposure control and content balancing in CAT.

2.2.2.1 Exposure control

In CAT, selecting items without considerations other than the objective selection criterion usually leads to a disproportionate use of particular items in the pool. That is, some items are much more frequently administered to examinees, and other items are rarely or never administered. Test developers do not want examinees to have pre-knowledge of the items and do not want to waste the cost of developing the unused items. To limit the exposure of items in CAT, exposure control procedures have been introduced by putting some constraints on item selection during the CAT administration (e.g., Chang & Ying, 1999; Davey & Parshall, 1995; Kingsbury & Zara, 1989; McBride & Martin, 1983; Stocking, 1993; Revuelta & Ponsoda, 1998; Sympson & Hetter, 1985). These exposure control procedures can be divided into four main types: randomized, conditional, stratified, and combined procedures (Georgiadou, Triantafillou, & Economides, 2007).

Randomized procedures include several variations on randomization of items in item selection for exposure control (e.g., Bergstrom, Lunz, & Gershon, 1992; Eignor et al., 1993;

Way, Zara, & Leahy, 1996). For instance, McBride and Martin's (1983) 5-4-3-2-1 procedure randomly selects the first item from a group of the most informative five at the beginning of the test. After the current proficiency is updated, a group of the four most optimal items are selected and the second item is chosen at random from this subset. This procedure continues until the subset is defined as the best single available item. Kingsbury and Zara (1989) proposed the randomesque procedure, which is the most commonly used in operational settings due to its simplicity. This procedure randomly selects one item from the most informative *n* items (e.g., five or seven) based on an examinee's current proficiency estimate throughout the entire CAT process.

Conditional procedures control the exposure of items based on a given criteria (e.g., the frequency of item usage for a target sample of examinees). The most representative example of conditional procedures is the Sympson-Hetter method (Sympson & Hetter, 1985). This procedure requires an item exposure parameter, say *k*, ranging from 0 to 1 (i.e., the conditional probability that the item will be administered given the item has been selected) obtained iteratively from simulations for a target sample of simulated examinees prior to the administration of CAT. The value of *k* is high for a certain item, indicating this item has been rarely administered and thus has a higher probability of being administered if selected. The value of *k* is low for a particular item, implying the item has been frequently administered and has a lower probability of being administered if selected. During the CAT administration, after selecting an optimal item to be administered, a random number from a uniform distribution between 0 and 1 is generated and compared to the exposure parameter *k* of the selected optimal item. This item is administered if this random value is smaller than the value *k* of the selected item. Otherwise, the next optimal item is selected, and the same procedure is applied to this item

until an item is administered to the examinee. This procedure successfully controls the over-exposed items, but it is very time-consuming because the iterative simulations must be done a priori (Georgiadou et al., 2007).

Stratified procedures stratify the item pool according to statistical properties such as item discrimination and difficulty, and then administer an item from a given stratum. The *a*-stratified method (Chang & Ying, 1999) is an example of the stratified methods. This procedure is motivated by the situation where items are solely chosen based on their information, resulting in disproportionate usage of some highly informative items. As informative items are unnecessarily used earlier in the test, in which the interim proficiency estimates contain too much error to be considered accurate, the final proficiency estimates are more likely to be over or under estimated. To regulate the use of highly informative items, this method first administers items with lower *a*-parameters at the earlier stages of the test and administers items with higher *a*-parameters at the later of the test to improve the efficacy of the items. Following this solution, many variations have been proposed, including the *a*-stratified with *b*-blocking (Chang, Qian, & Ying, 2001) and the 0-1 stratification strategy (Chang & van der Linden, 2003), etc.

Lastly, combined procedures attempt to combine two or more exposure control methods. Revuelta and Ponsoda's (1998) progressive-restricted combined procedure is a notable example. This combined procedure, derived from the maximum information method and the restricted maximum information method, is intended to prevent the overexposure of items and to increase the usage of rarely or unused items while maintaining precision of proficiency estimates. The modified version of this method, the progressive-restricted standard error method was also developed (see McClarty, Sperling, & Dodd, 2006 for details).

Among these exposure control procedures, I choose the randomesque method, the Sympson-Hetter method, and the *a*-stratified with *b*-blocking method to see how the different procedures affect the level of customization for CAT using the proposed adaptivity statistics (see Section 4.4.1.2).

2.2.2.2 Content balancing

Like the P&P test, a CAT should conform to a test blueprint, especially to cover multiple content areas, which is closely associated with the interpretation and validity of the test scores. This can be realized through content balancing procedures. Although a variety of strategies for content balancing exist, the most commonly used procedure in research and operational settings is Kingsbury and Zara's (1989) constrained CAT. In this procedure, the target proportions of each content area are first prespecified. After the administration of each item, the current proportions of each content area are calculated and compared to the pre-specified target proportions. The content area with the largest discrepancy between the target and current proportions is selected, while items from other content areas are filtered out from the item pool, and the next item with the highest information will be selected from the available items from that content area. Previous research (e.g., McClarty et al., 2006) has provided evidence to support that this procedure successfully administers specified proportions of items per content area.

In addition to this simple procedure, more complex strategies for content balancing are also available. These content balancing methods include the weighted deviations model (Swanson & Stocking, 1993), the shadow test approach (van der Linden & Reese ,1998), the weighted penalty model (Shin, Chien, Way, & Swanson, 2009), the maximum priority index method (Cheng & Chang, 2009), and the bin-structured method (Davey, 2005), among others.

2.2.3 Scoring procedure

In the beginning of the test, an initial proficiency value is arbitrarily determined because there is no available information about an examinee. The initial proficiency value is typically set to 0.0, which is the mean of the proficiency's distribution in the test population. The proficiency estimate is then updated after each item is administered based on the item responses. Proficiency estimation methods are essential because the methods could affect not only the reporting score of the test, but also the selection of items to be administered and the decision of terminating the test (e.g., standard error of proficiency estimates). Previous studies have proposed proficiency estimation approaches (e.g., Bock & Mislevy, 1982; Lord, 1986; Owen, 1975), provided ways to overcome some challenges that a particular estimation method has for CAT (e.g., Han, 2016) and compared their performance, as well (e.g., Wang & Vispoel, 1998). Among the existing proficiency estimation methods, maximum likelihood estimation (MLE) and Bayesian estimation methods such as expected a posteriori (EAP; Bock & Mislevy, 1982), maximum a posteriori (MAP; Samajima, 1969), and Owen's empirical Bayesian method (Owen, 1975) are the most widely used in CAT programs.

MLE determines the most likely location of an examinee's proficiency by multiplying the probabilities of a response string with the location independence assumption (Hambleton & Swaminathan, 1985). To find the most likely value of proficiency estimates that maximizes the likelihood, the Newton-Raphson method can be used. The MLE approach provides proficiency estimates which are consistent, efficient, and asymptotically normally distributed. The normality property is a very practical advantage of MLE because it allows the standard error of the proficiency estimate to be calculated using the information function shown in Equation (2.4). However, the MLE provides an infinite proficiency estimate if the item responses are either all

correct or incorrect so that at the beginning of CAT, the estimates cannot be computed until both correct and incorrect responses exist. To tackle this problem, in practice, either a step parameter (e.g., 0.7; Reckase, 1976) or arbitrary lower and upper bound values of proficiency estimates (e.g., say -4 and +4) are used early in the CAT. Another way to solve this issue is to start with a Bayesian estimation procedure and switch to MLE after both correct and incorrect responses are obtained (e.g., NCSBN, 2016).

Bayesian estimation methods are alternatives to MLE for handling this infinity problem. EAP determines the most likely location of proficiency as the expected mean of the posterior distribution, and MAP as the model of the posterior distribution. These Bayesian approaches can estimate the examinee's proficiency level even after the first response is obtained with the help of the prior distribution. Although the Bayesian estimation methods have such an advantage, a well-known weakness is that their estimates are generally biased toward the mean of the prior distribution, resulting in a shrunken score scale (e.g., Ho & Dodd, 2012; Kim & Nicewander, 1993; Wang & Vispoel, 1998; Weiss, 1982). Another example of a Bayesian estimation approach is Owen's empirical Bayesian method. At every update of proficiency estimate in CAT, the posterior proficiency distribution from the previous one is used as a prior distribution for the estimation. Owen's Bayesian method is also very popular because it is straightforward to compute the proficiency estimates and faster than other Bayesian methods. However, this method has the major downside that the proficiency estimates are affected by the sequence of the item presentation. This problem might be alleviated by re-estimating the response strings at the end of CAT using an alternative proficiency estimation method such as maximum likelihood estimation (Wang & Vispoel, 1998).

Taken together, I focus on MLE and EAP (Section 4.2.2.3). These two proficiency estimation methods are notably used in CAT (Hambleton, Swaminathan, & Rogers, 1991, p. 148; Weiss, 1982).

2.2.4 Stopping rules

Stopping rules are closely tied with the purpose of tests. In general, there are two ways to decide when the test terminates: fixed length of the tests and variable length of the tests. A fixed-length test requires all the examinees to receive an equal number of items given in the test. However, this feature of the same number of items that every student takes might cause the measurement precision of final proficiency estimates to differ across students' proficiency levels depending on the distribution of items in the pool.

A variable-length test provides a different number of the items to students until a pre-specified standard error (i.e., measurement precision) of proficiency estimates is satisfied. A target measurement precision (e.g., < 0.3 or 0.2) is considered a test termination criterion in order for each examinee to have the same magnitude of measurement precision. One problem in variable-length CAT is that the examinees with very high or low proficiency levels will have a longer test than others due to the fact that the item pool could run out of suitable items to be administered. One suggested approach to deal with this issue is to combine the measurement precision rule with setting the maximum/minimum number of items in practice (Thissen & Mislevy, 2000). In this dissertation, all simulation studies of CAT were based on the fixed-length test, and the empirical illustration using an operational adaptive test was a variable-length CAT.

2.2.5 Adaptive test designs

Due to the benefits drawn from CAT, there has increased the applications of CAT with some modifications in test designs for compensating for its weaknesses and for encouraging the

practical uses in real educational and operational settings. For example, the full item-level CAT cannot review in advance the test items to be administered to each examinee, implying the potential of a lack of quality control (Luecht & Nungester, 1998). Also, the full CAT may require more funding for its development and implementation.

However, multistage adaptive testing (MST), as a special form of CAT, adapts at the stage/module level, and it has some practical advantages over the item-level CAT in operational settings (e.g., Stark & Chernyshenko, 2006). With MST, examinees can not only skip items but also review and revise their responses to the items within the stage during the testing, which is not available in CAT. Modules (i.e., a group of items) are also pre-assembled before test administration. So, MST allows test developers to control the quality of tests and content balancing while maintaining a comparable measurement precision to the full CAT when the test is well designed (Xing & Hambleton, 2004). However, MST may reduce adaptivity compared to the item-level adaptation of CAT (e.g., Reckase et al., 2019).

Recently, another form of CAT, called hybrid CAT (Wang, Lin, Chang, & Douglas, 2016), has been introduced that combines characteristics of item-level CAT and MST. Administering an MST in the beginning of the test contributes to improving an initial proficiency estimate for the later implementation of CAT but also achieving content balancing more systematically. In this dissertation, how much adaptation occurs across proficiency levels is examined depending on the different adaptive test designs of the item-level CAT and MST (see Section 4.5.2).

2.3 Factors Affecting Adaptation

The amount of adaptation can be affected by numerous factors associated with the characteristics of an item pool and the CAT specifications. First of all, the item pool is a

fundamental and vital element for the development and deployment of a CAT. The best, sophisticated CAT program cannot function well if an item pool consists of poor-quality items or items suitable for the limited range of proficiency (Flaugher, 2000; van der Linden, Ariel, & Veldkamp, 2006). The higher the quality of the item pool, the more likely the adaptive algorithm will work well. Accordingly, it needs to understand the extent to which the amount of adaptation during the CAT would be affected by characteristics of the item pool. To do this, previous studies examined the effects of the item pool's characteristics on the adaptation in terms of item pool size and the spread of difficulty of the items in the pool at the entire group level (Ju & Lee, 2018; Kim, Ju, & Reckase, 2018; Reckase et al., 2018). The results of these studies suggested that the item-pool composition would, in predictable ways, influence the amount of adaptation. That is, the item pool should be about more than ten times the test length with more spread in difficulty of items for the adequate adaptivity of the CAT. In addition to that, the shape of the item pool could affect the results of the CAT and, plausibly, the performance of the adaptive algorithm, taking into account the shape of the proficiency distribution (e.g., Gönülates, 2015; Reckase, 2010).

Intertwined with the item-pool characteristics, a variety of components of the CAT would also impact the consequences of the CAT, including adaptivity. For example, IRT models would affect the measures of adaptivity such as the IRT model's scaling and scoring functions, which may eventually affect the selection of items at the momentary proficiency estimate. Kim et al. (2018) compared the overall adaptation measures using the 3PL model to those using the Rasch model. Because of the effects of discrimination and guessing parameters, the suggested benchmark values were slightly different for the two models, though their conclusions appeared to be the same.

Meanwhile, proficiency estimation plays a pivotal role in a successful CAT implementation because it is closely related to the item-selection procedure. The MFI item selection method is the most frequently used in the CAT because of measurement precision and efficiency. This method assumes a perfect correspondence between the current proficiency estimate and the true proficiency level of an examinee. If the assumption is violated due to poor accuracy of the proficiency estimates, the item-selection algorithm may select items that are not well associated with the target true proficiency, resulting in selecting less optimal items that contribute to increasing errors in subsequent proficiency estimates (Ho & Dodd, 2012). This issue might be more severe in the early stages of CAT because, generally speaking, little adaptation occurs before the proficiency is well estimated. Given this fact, in previous research (Reckase et al., 2018), adaptation statistics were computed using all of the items administered to examinees during a CAT and also for the items used in the last half of their tests. It was shown that higher values of the adaptivity statistics were reported with the last half of the test, though the extent of the increment was small.

Most previous research has used MLE in the CAT. Ju and Lee (2018) explored the performance of the overall adaptivity measures across different proficiency estimators. They found that the correlation index and the ratio of standard deviations index were robust to different estimation methods. Yet the proportion of reduction in variance (PRV) index appeared to be affected by the proficiency estimates, especially with an increase in the PRV benchmark value using the EAP method. This impact was due to the Bayesian estimator's property of regressing toward the mean or mode of the prior distribution (e.g., Kim & Nicewander, 1993). Test length also might matter; after all, with a longer test, there might be more chances of a test

being customized for a test taker. However, the adaptivity measures appeared to be robust to the test length (Ju & Lee, 2018).

Constraints on the item-selection algorithm should negatively influence the selection of optimal items at the interim proficiency estimate during a CAT, resulting in a degrading of the amount of adaptation. Constraints include content constraints, exposure-control constraints, and item-type constraints. Previous research (Reckase et al., 2017, 2018) has demonstrated the effects of exposure control on adaptation. For example, the Sympson-Hetter exposure-control procedure seemed to limit the amount of adaptation with a relatively small item pool; no limit, however, was shown with the randomesque procedure and *a*-stratified with *b*-blocking procedure. Recently, many content constraints can be easily controlled through, among others, constrained CAT, shadow test approach, weighted deviation model.

Observations of operational CATs have presented that all test designs are not equally adaptive because of different units of the customization of the test and designs of test specifications. For instance, a full regular CAT adapts at the item level, while MST adapts at the stage/module level. In recent years, researchers have introduced a new hybrid design that incorporates both item-level and stage-level CAT (Wang et al., 2016). Reckase and colleagues (2017; 2019) compared the adaptivity across-item level CAT and different designs of MST, identifying that the MST design appeared to be less adaptive than the others.

Taken all together, this dissertation explores, among the various factors that affect the amount of adaptation during a CAT, the interaction effects of the pool characteristics and proficiency estimators on a new class of conditional adaptation indices. The effects of constraints for exposure control and adaptive testing designs on the amount of adaption are additionally examined through simulations.

CHAPTER 3.

INDICES FOR THE AMOUNT OF ADAPTATION

Before discussing the indices to measure the amount of adaptation, it is necessary to define operationally what test adaptation (i.e. adaptivity) is. Reckase and his colleagues (Reckase et al., 2018, 2019) defined adaptation as the extent to which a CAT gives items that properly match the final proficiency estimate for the examinee. Kingsbury and Wise (2018) similarly defined test adaptation but with more focus on test information, given the available item pool and test specifications. In this dissertation, test adaptation or adaptivity can be defined as the extent to which a CAT provides the informative items that properly match current proficiency estimates at each stage of the CAT process. Thus, a test can be viewed as being highly adaptive when the items administered to each examinee match well with the provisional proficiency estimates at the start of each item during the CAT.

To quantify the amount of adaptation of a CAT, it is assumed that test taker *j* has a known location on a latent continuum (θ_j) and that the goal of the CAT is to select the optimal set of items that will produce an accurate estimate of that location ($\hat{\theta}_j$) given the available item pool and CAT specifications (Reckase et al., 2018). In the hypothetical case in which the location of the test taker on the continuum is known with an infinite item pool, an optimal set of test items for each test taker *j* would have maximum information at θ_j when the maximum Fisher information (MFI) item selection method is used.

Consider the simple case of the Rasch model. In that hypothetical case, all the selected items would have item-difficulty parameters (i.e., *b*-parameters) equal to θ_j, resulting in a set of items that had the average of *b*-parameters equal to θ_j and their standard deviation equal to zero.

Extending this case to a sample of test takers with true locations on the proficiency scale, the mean *b*-parameter for each examinee would be perfectly correlated with θ_j, and the standard deviation of the mean *b*-parameters would be equal to that of θ_j's (Reckase et al., 2018).

Alternatively, using the 3PL IRT model for scaling and scoring, the location of the maximum information for each optimal item *i* for a test (θ_i^*; Birnbaum, 1968) can be substituted for the *b*-parameter:

$$\theta_i^* = b_i + \frac{1}{Da_i}\log\left(\frac{1+\sqrt{1+8c_i}}{2}\right) \tag{3.1}$$

where a_i is an item-discrimination parameter, b_i is an item-difficulty parameter, c_i is an item-pseudo-guessing parameter, and *D* is a scaling constant that makes the logistic function similar to the normal ogive function. Since the location (θ_i^*) of maximum information is slightly higher than the *b*-parameter, the selection of items might be a little bit different than the selection based on the difficulty parameter, but the concept of the adaptation is still the same with use of the location (θ_i^*) of maximum information (Kim, Ju, & Reckase, 2018).

This ideal type of CAT never exists in real operational settings because we never know the true location of a test taker on the proficiency continuum. Nevertheless, the hypothetical case does give some direction toward the possible types of measures that can be used to quantify the amount of adaptation that occurs in an operational adaptive testing. This conceptualization of adaptation and the ideal features of a CAT thus leads to existing adaptivity measures (Reckase et al., 2018; 2019) and a new class of conditional statistics of the amount of adaptation proposed in this dissertation. This chapter reviews the current measures of the amount of adaptation and then introduces three new conditional adaptation indices.

3.1 Existing Measures of the Amount of Adaptation

Under the conceptualization and assumptions of a desired CAT, Reckase and his colleagues (Reckase et al., 2018) proposed three overall statistics: Correlation index, ratio of standard deviations index, and proportion of reduction in variance index. These measures were mostly based on the variance of the *b*-parameters (or the location of maximum information) for the items administered to test takers. Note that the location of maximum information in Equation (3.1) can be substituted for the *b*-parameters for computing the three statistics when the 3PL model is used. They performed well in assessing the overall adaptivity of a CAT over examinees; they could not, however, be applied to evaluate the adaptation contingent on proficiency level. Motivated by this perceived concern, Kingsbury and Wise (2018) introduced a new index of the amount of adaptation using the IRT test information that could be used to diagnose adaptivity for both the entire group and the individual test events.

3.1.1 Correlation index

The first adaptivity measure that Reckase and his colleagues proposed is the *correlation* between the mean *b*-parameter for the items administered to examinees and the final estimate of their proficiency:

$$Correlation\ Index = r(\bar{b}_j, \hat{\theta}_j) \tag{3.2}$$

where $\bar{b}_j$ is the mean *b*-parameter for the items administered to a test taker *j*, and $\hat{\theta}_j$ is the final estimate of the location on the θ-scale for a test taker *j*. This index indicates whether examinees with various levels of proficiency receive tests that are different in difficulty and that the difficulty levels match well the estimated proficiency levels. As shown in Reckase et al. (2018), higher values of the index imply better adaptivity of a CAT. The suggested benchmark value for

interpreting this statistic is for the Rasch model the “low .90s” (Reckase et al., 2018) and for the 3PL model the “high .90s” (Kim et al., 2018).

3.1.2 Ratio of standard deviations index

Even if the correlation index shows a high value close to 1.0, it is possible that the adaptivity of the CAT might not be good because of poor qualities of the item pool or some problems with the item selection algorithm. The second index helps assess such aspects of adaptivity. It is the *ratio* of the standard deviation of the averages of the *b*-parameters for the items administered to examinees, $s_{\bar{b}_j}$, to the standard deviation of the final proficiency estimates for those examinees, $s_{\hat{\theta}_j}$:

$$Ratio\ of\ SD\ Index = s_{\bar{b}_j}/s_{\hat{\theta}_j} \tag{3.3}$$

where the subscript *j* indicates the particular examinee. This index indicates whether the spread of the mean *b*-parameters of the items selected to examinees matches the spread of their proficiency estimates. If the item selection algorithm is working properly but an item pool has a limited range of difficulty, the correlation index may yield a high value, but this ratio index may report a lower value because of the small $s_{\bar{b}_j}$ relative to $s_{\hat{\theta}_j}$ (Reckase et al., 2018).

For this statistic, unlike other adaptation indices, the value of 1.0 is optimal, as higher values than 1.0 can be obtained. For instance, values larger than 1.0 can be obtained when the item pool has an unusual distribution of the *b*-parameters with many extremely easy and difficult items but insufficient items in the middle range of difficulty. In this case, the $s_{\bar{b}_j}$ value could be large relative to $s_{\hat{\theta}_j}$, ending up with the index value greater than 1.0. Therefore, the distance from 1.0 is important when interpreting this statistic for evaluating adaptivity. Since the unusual type of item pool is rarely found in the real word, previous studies suggested the benchmark

value below 1.0, which for the Rasch model is the "middle .80s" (Reckase et al., 2018) and the "high .70s" for the 3PL model (Kim et al., 2018).

3.1.3 Proportion of reduction in variance index

The last index that Reckase et al. (2018) introduced is the *proportion of reduction of the variance* (PRV) of the *b*-parameters for the items selected for the examinee, on average, from the amount of variance of the *b*-parameters for all of the items in the pool:

$$PRV = \frac{s_b^2 - pooled\ s_{b_j}^2}{s_b^2} \tag{3.4}$$

where $pooled\ s_{b_j}^2$ is the average of the within-examinee variances of *b*-parameters for the items selected for each examinee, and s_b^2 is the variance of the *b*-parameters for all the items in the pool. This index focuses more on the adaptivity within the examinee regarding the item pool, especially in a situation where the item pool has insufficient items in the area in which the final estimate of the examinee's proficiency is located. If such a situation is encountered, the adaptation of the CAT may also be poor because the item selection algorithm may have to select items whose *b*-parameters poorly match the current proficiency estimates. Hence, the variation of the *b*-parameters for that test taker might be large, though the mean *b*-parameter might be close to the final proficiency estimate. The index would reflect this situation and be constructed in the same form as Hoyt's reliability (Hoyt, 1914).

Regarding the interpretation of the PRV indicator, a value less than 1.0 represents the average amount of within-examinee variation in difficulty of the items administered over a sample of examinees relative to the amount of variation in difficulty for all the items in the pool. That is, if the variation of *b*-parameters is zero – meaning that for each examinee it is constant as in the aforementioned hypothetical ideal case – but the item pool has variation in *b*-parameters

for the items, then this PRV value is 1.0. The suggested benchmark value is .80 regardless of the IRT model (Kim et al., 2018; Reckase et al, 2018).

3.1.4 Percent of optimal information index

Kingsbury and Wise (2018) introduced a new statistical indicator to measure the amount of adaption using the IRT test information that can be applied to not only a group of examinees but also to individual or subgroup test events. Their index, called the *percent of optimal information* (POI), is based on the ratio of observed test information to the maximum information possible given the item pool and the IRT model and defined as follows:

$$POI = 100 * \frac{\sum \frac{TA_j}{TO_j}}{J} \tag{3.5}$$

where TA_j is the actual test information observed for an examinee *j* based on its final estimated proficiency, and TO_j is the optimal amount of test information that can be obtained by administering a 40-item test at the true proficiency level of an examinee from a given item pool and IRT model. An alternative way to compute the optimal information is to calculate the test information available in the pool from the most informative test items at the final estimated proficiency. By the summation over the examinees in the group of interest (i.e., *J* refers to the group size), the POI index can also be used as an overall measure of the adaptation. This index is easily interpretable, but it has not yet been thoroughly examined across numerous item pool conditions or constraints on item selection. In addition, it is still blind to the extent information obtained during the CAT process based on interim proficiency estimates.

3.2 New Conditional Measures of the Amount of Adaptation

This dissertation proposes new three indices to investigate the amount of adaptation conditional on an examinee's proficiency level using (1) *b*-parameters of items administered to each examinee (deviation of difficulty; DOD), (2) their within-examinee variances (conditional proportion of reduction in variance; CPRV), and (3) the IRT information (ratio of information; ROI). These statistics have the same assumptions and the same goal of the CAT to the overall adaptation measures (Reckase et al., 2018). The only difference is that these new measures can evaluate the various aspects of adaptivity that result from the implementation of the CAT conditional on the proficiency level or by subgroups of test events. Also, they focus more on the characteristics of the items based on the *interim* proficiency estimates during the CAT process, instead of the final proficiency estimates at the end of the tests. Note that like the overall adaptation statistics described above, the *b*-parameters can be replaced with the location of maximum information (Birnbaum, 1968) when the 3PL model is used for scaling and scoring.

3.2.1 Deviation of difficulty index

The first index that the study proposes is the *deviation of difficulty* (DOD) index that focuses on the observed difference between the *b*-parameter of the administered item and the examinee's interim proficiency estimate at which that item was selected (i.e., desired item difficulty). The DOD index can assess how well a CAT uses the available item pool to match item characteristics to the examinee's provisional proficiency estimate. It can also allude to how well the potential efficiency of an item is realized, given the fact that the expected efficiency gain is attained if the examinee's proficiency is close to the location of that item.

The DOD index for each examinee *j* is, over the test items administered, the average proportionate reduction of the observed location match between the item and the interim

proficiency estimate relative to the average deviation of all the eligible items in the pool from the current proficiency estimate. The index is represented by:

$$DOD_j = \frac{1}{L_j - n_i} \sum_{i=ni+1}^{L_j} \frac{\frac{1}{PS}\sum_{h=1}^{PS}\left|b_h - \hat{\theta}_{(i-1)j}\right| - \left|b_{ij} - \hat{\theta}_{(i-1)j}\right|}{\frac{1}{PS}\sum_{h=1}^{PS}\left|b_h - \hat{\theta}_{(i-1)j}\right|} \tag{3.6}$$

where $\hat{\theta}_{(i-1)j}$ is examinee j's interim proficiency estimate prior to selecting the *i*th item, b_{ij} is the difficulty parameter of the *i*th item for the examinee *j*, L_j is the test length for examinee *j*, and PS is the number of available items in the pool at the interim proficiency estimate, $\hat{\theta}_{(i-1)j}$. Note that n_i is the number of initial items before the first update of the proficiency estimate occurs. For instance, for fully adaptive testing, a single initial item is generally administered, while for multistage adaptive testing, a group of items in the routing module may be administered. Since there is no interim proficiency estimate other than the arbitrary starting value prior to selecting the first item(s), the initial item set is not taken into account in the index calculation.

The DOD index is a concept similar to that of the examinee *j*'s difficulty mismatch (DM) index to quantify the CAT's difficulty alignment (Wise, Kingsbury, & Webb, 2015):

$$DM_j = \frac{\sum_{i=2}^{L}|\hat{\theta}_{ij} - b_{ij}|}{(L-1)\sigma}$$

where σ is the standard deviation of proficiency level estimates. While useful with the similar interpretation of *z*-scores, this DM index does not have an upper limit of the possible values, and it considers only the distance of difficulty from the provisional proficiency estimate. It thus needs some criteria in advance to provide the upper limit indicating a high informative test.

However, the proposed DOD index is readily interpretable. The value of 1.0, the highest attainable, indicates a test event where items were perfectly matched to the momentary proficiency estimate at each item selection. The distance from 1.0 represents an average of the

deviation of the difficulty of the administered item from the interim proficiency estimate relative to the average deviation of the difficulty of all the eligible items in the pool from that proficiency estimate. Thus, a higher value of DOD_j indicates a higher match level, implying better adaptation in that the CAT is providing, throughout the test administration, close to the maximum information available at the interim proficiency estimate.

3.2.2 Conditional proportion of reduction in variance index

The second proposed index is the *conditional proportion of reduction in variance* (*CPRV*) index, which is a modified version of the PRV index. It determines if the item pool has sufficient items in the region of the final proficiency estimate of each examinee. This index would be particularly useful in a situation where the item selection algorithm may have to select some items whose difficulty parameters poorly match the current proficiency estimate. Hence, the variation of the difficulty parameters for that examinee might be large, though the mean difficulty might be close to the final proficiency estimate. The CPRV is expressed as:

$$CPRV_j = \frac{s_b^2 - s_{b_j}^2}{s_b^2} \tag{3.7}$$

where s_b^2 is the variance of the *b*-parameters for all the items in the item pool and $s_{b_j}^2$ is the within-examinee variance of the *b*-parameters for the items administered to examinee *j*. Like the PRV index, a value deviating from 1.0 indicates the amount of variation in difficulty for the items selected for examinee *j* compared to the amount of variation in difficulty for the items in the full item pool.

3.2.3 Ratio of information index

The last index, called the *ratio of information* (*ROI*) index, is equal to, over the administered L_j-items for examinee *j*, the average ratio of the information of item i_k at the

interim proficiency estimate prior to selecting the *k*th item *i* for examinee *j*, $I_{i_k}[\hat{\theta}_{(k-1)j}]$, to the maximum potential information that item *i* can have, $I_i[\hat{\theta}_i^*]$:

$$ROI_j = \frac{1}{L_j}\sum_{i=1}^{L_j}\frac{I_{i_k}[\hat{\theta}_{(k-1)j}]}{I_i[\hat{\theta}_i^*]} \tag{3.8}$$

where $\hat{\theta}_i^*$ is the point at which the item *i* can reach maximum information (Birnbaum, 1968; e.g., $\hat{\theta}_i^* = b_i$ for 1PL/2PL model). Alternatively, the observed information of each of the administered items can be computed at the final proficiency estimate, $\hat{\theta}_j$, rather than the interim estimate. While being readily useful in practice, this method could be blind to the appropriateness of the items customized to the examinee in the middle of the CAT administration. For instance, if a high-proficiency student, say θ = 2, begins with an easy item (due to an initial estimate of 0.0) but happens to miss that item, the student will get the low proficiency estimate after the first item, leading to the student receiving a couple of relatively easy items for the next few items; however, if the student then improves the proficiency estimate continuously by answering correctly all the rest of the items, the ROI value using the final proficiency estimate shows the test is relatively low informative because the student gets less informative items close to the examinee's final proficiency estimate. The original ROI index in Equation (3.8), on contrary, indicates good informative items presented to the student during the CAT process. Overall, this index can assess how informative a test is compared to the maximum potential information that the administered items can have. The ROI index can range from 0.0 to 1.0. A value of 1.0 indicates that a test is appropriately constructed and administered to the examinee using the items whose maximum potential information is realized at that examinee's proficiency. The higher the values, the more informative a test for the examinee.

In addition to the utilization of ROI conditional on the proficiency level in Equation (3.8), it can also be used for the overall diagnosis of adaptivity by simply averaging the ROI values over the entire group of examinees:

$$ROI_j = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{L_j}\sum_{i=1}^{L_j}\frac{I_{i_k}[\hat{\theta}_{(k-1)j}]}{I_i[\hat{\theta}_i^*]}\right) \tag{3.9}$$

where *N* is the total number of examinees that took the adaptive test. It would help test developers or practitioners understand the overall picture of adaptivity of CATs at the entire group or target group levels of interest.

The ROI index is originally derived from the concept of relative efficiency (Lord & Novick, 1968) that compares the Fisher information functions. This may also be in line with the POI index (Kingsbury & Wise, 2018). However, the ROI index is conceptualized differently from the POI index with respect to the definition of optimal information (i.e., the denominator of the index). Kingsbury and Wise (2018) identified the optimal test information through the administration of the entire test at the true or final proficiency level using the actual item pool. They also stated that the optimal information can be defined using the theoretical limit from the known value of the maximum information given the Rasch model. This is actually a similar concept to the item-pool utilization index (Gönülates, 2015) used for evaluating the efficiency of item pool performance. For the proposed ROI index, however, the optimal information focuses more on the maximum potential that the administered item has; this is similar to the expected item efficiency that Han (2012) used as a step to item selection. Thus, the ROI index is expected to evaluate the adaptivity of the CAT from whether an administered item fulfills the maximum level of the attainable information at the examinee's interim proficiency estimate. Moreover, the ROI index is reflective of how well the items are informatively presented to the examinees

during the whole process of the CAT, while the POI index cares more about whether informative items are provided to the examinee around the final proficiency estimate.

CHAPTER 4.

METHODS

This dissertation proceeds with five main studies to evaluate the feasibility and utility of three new conditional indices to measure the amount of adaptation in the implementation of computerized adaptive testing (CAT) with dichotomously scored items using simulated data and real operational CAT data. All the analyses were conducted using MATLAB R2015b (The MathWorks, Inc., 1984-2015) and the visualization of the results were completed using R software (R Core Team, 2018). This chapter describes details about the research designs for replying to the five research questions (see Section 1.2).

4.1 Common CAT Specifications

All item-level CATs in the first four studies share some common CAT specifications. The CAT was a fixed-length test of 40 items. An initial item-level proficiency estimate of 0.0 was used for all examinees. Items were selected using the maximum Fisher information (MFI) algorithm that chooses the item to be administered that has the maximum information at the current proficiency estimate. Other than Research Question 1, maximum likelihood estimation (MLE) was used to estimate the interim and final proficiency after both correct and incorrect responses existed in the response string. When only either correct or incorrect responses are present, the maximum likelihood estimates are infinite. To deal with this problem, prior to MLE, the last proficiency estimate was incremented by the step size of 0.7 after a correct response and decremented by 0.7 after an incorrect response (Patience & Reckase, 1980; Reckase, 1975). Also, maximum likelihood estimates were confined between -4 and 4 to restrict some extreme proficiency estimates within a practical interval.

For each study condition, item-level CATs were administered to 2,000 examinees randomly sampled from a standard normal distribution, N (0, 1). This sample size is reflective of large-scale operational testing settings to get a representative sample from the proficiency population. To simulate examinee's responses, a random number was drawn from the uniform distribution ranging from 0.0 to 1.0. The random uniform number was then compared to the examinee's probability of correctly answering the item to determine the examinee's response for the item. If the probability of correct response was greater than the random number, a score of 1 was assigned as a response; otherwise, a score of 0 was recorded. In all cases, the results were replicated 50 times for computing the stability of the adaptivity statistics.

4.2 Research Question 1

The first set of simulations were intended to evaluate the sensitivity of the three conditional adaptation indices to various item-pool quality conditions, proficiency estimators, and IRT models with the goal of providing some guidelines for interpreting these indices. It was hypothesized that the values of the conditional adaptivity indices will increase as the item pool size and item pool spread increase, but that the indices will be rarely affected by proficiency-estimation methods and IRT models.

4.2.1 Item pool

To simulate an item pool for the 3PL model as realistically as possible, the simulated item pool was modeled after the multivariate distribution of the *a*-, *b*-, and *c*-parameters but with different marginal distributions, respectively, using only multiple-choice items in the item pool from the Minnesota Comprehensive Assessment (MCA) Grade 6 Mathematics adaptive assessment. The descriptive statistics and zero-order Pearson correlations of the item parameters are presented in Table 4.1. Specifically, while taking into account the correlation among the item

parameters, *a*-parameters were drawn from a lognormal distribution, *b*-parameters from a normal distribution, and *c*-parameters from a beta distribution. Based on the multivariate distribution, sets of item pools that act like the empirical pool were generated according to simulation conditions of interest. Note that an item pool based on the Rasch model was generated, only taking into account the distribution of *b*-parameters shown in Table 4.1.

Table 4.1

Descriptive Statistics and Zero-Order Correlations of Item Parameters for the Item Pool from Minnesota Comprehensive Assessment (MCA) Grade 6 Mathematics Adaptive Test (n = 635)

	Descriptive Statistics				Correlations		
	M	*SD*	Min.	Max.	*a*-parameter	*b*-parameter	*c*-parameter
a-parameter	1.03	0.30	0.20	1.99			
b-parameter	0.27	0.95	-2.53	3.14	0.24		
c-parameter	0.16	0.10	0.00	0.60	0.06	0.00	

4.2.2 Simulation design

The first simulation study was conducted to examine the sensitivity of the three conditional indices to item pool characteristics, proficiency estimator, and IRT model. Here, the item-pool quality was operationalized by two aspects: (1) item-pool size and (2) item-pool spread in *b*-parameters. These item-pool characteristics were fully crossed with proficiency estimators and IRT models, forming a total of 72 conditions (10 pool sizes × 2 proficiency estimators × 2 IRT models + 8 pool spreads × 2 proficiency estimators × 2 IRT models).

4.2.2.1 Item pool size

Using each IRT model, 10 item pools were generated that varied in item pool size from 50 to 500 in increments of 50. First, using the observed multivariate distribution described above, 500 sets of item parameters were generated that had descriptive statistics and correlations

as similar as possible to the empirical set in Table 4.1. These full sets of item parameters were then randomly divided into 10 sets with 50 items each in a way that item characteristics were similar across the 10 sets. Then, the first set of 50 items were used for the simulation of a 50-item pool. The first set was then combined with the second set of 50 items to construct the 100-item pool. This process was repeated, adding a set of 50 items each time, until the simulation was conducted using the full set of 500 items in the pool. This elaborated way of creating different sizes of item pools allows a researcher to solely explore the relationships between pool size and values of the adaptivity statistics. Otherwise, it is possible that a small item pool with high-quality items (i.e., items with high discrimination and small guessing parameters) could perform better than a larger pool with low-quality items given no other constraints imposed on item selection. Table 4.2 summarizes descriptive statistics and correlations among item parameters for the 10 generated item pools.

4.2.2.2 Item pool spread

Another aspect of item-pool characteristics that could affect the amount of adaptation is the degree of spread in difficulty of the items in the pool. That is, if the difficulty of items is in a limited range, even with the large item pool, the adaptive test cannot be suitably customized for examinees who are located outside that range. To quantify this situation, a CAT was simulated using eight 400-item pools that differed in the level of standard deviation of *b*-parameters from 0.2 to 1.6 in increments of 0.2 with the mean of *b*-parameters set to 0.0. Other *a*- and *c*-parameters were controlled to be the same as the 400-item pool generated above and were also fixed across all the item pools. Table 4.3 displays the summary of the simulated item pools by the level of spread in *b*-parameter. In all the conditions manipulated by the pool spread, the item

pool size was 400, which is at least 10 times larger than the test length of 40 items as recommended by Stocking (1994).

4.2.2.3 Proficiency estimation methods

Two proficiency estimation methods were considered. One was MLE, most frequently used in operational settings, and the other was expected a posteriori (EAP; Bock & Mislevy, 1982) using the standard normal distribution as the prior distribution and using 81 evenly spaced quadrature points to determine the posterior distribution.

4.2.2.4 IRT models

The performance of the three conditional indices were further inspected using two IRT models: (1) Rasch model (i.e., one-parameter logistic model) and (2) three-parameter logistic (3PL) model. Comparing the performance between the two IRT models can inform us of how the *c*-parameter affect the stability of the indices over the proficiency continuum.

Table 4.2

Descriptive Statistics of Generated Item Pools by Item Pool Size

Pool Size	*a*-parameter				*b*-parameter				*c*-parameter				Correlation		
	M	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	(*a*, *b*)	(*a*, *c*)	(*b*, *c*)
50	0.98	0.27	0.57	1.91	0.10	1.00	-1.85	1.89	0.17	0.09	0.01	0.54	.24	.00	.06
100	1.00	0.26	0.57	1.91	0.28	0.94	-1.85	2.29	0.16	0.10	0.01	0.59	.27	.00	-.03
150	1.01	0.27	0.54	1.95	0.33	1.00	-2.43	2.96	0.16	0.11	0.01	0.59	.24	.05	-.01
200	0.99	0.27	0.54	1.95	0.31	0.98	-2.43	2.96	0.16	0.10	0.01	0.59	.24	.08	-.00
250	1.01	0.29	0.54	2.07	0.28	0.95	-2.43	2.96	0.16	0.10	0.01	0.59	.23	.01	.02
300	1.01	0.29	0.52	2.23	0.27	0.96	-2.43	2.96	0.16	0.10	0.01	0.59	.26	.04	.00
350	1.02	0.30	0.52	2.23	0.28	0.96	-2.43	2.96	0.16	0.10	0.01	0.59	.28	.05	.04
400	1.02	0.30	0.52	2.23	0.27	0.96	-2.43	2.96	0.16	0.10	0.01	0.59	.24	.07	.03
450	1.02	0.29	0.46	2.23	0.27	0.96	-2.43	3.19	0.16	0.10	0.01	0.59	.26	.06	.00
500	1.03	0.30	0.46	2.35	0.27	0.95	-2.43	3.19	0.16	0.10	0.01	0.60	.24	.03	-.01

Note. The simulated item pool based on the Rasch model had the same distribution of *b*-parameters.

Table 4.3

Descriptive Statistics of Generated Item Pools by Item Pool Spread (n = 400)

SD of *b*-parameter	*a*-parameter				*b*-parameter				*c*-parameter				Correlation		
	M	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	(*a*, *b*)	(*a*, *c*)	(*b*, *c*)
0.2	1.02	0.30	0.52	2.23	0.00	0.20	-0.55	0.54	0.16	0.10	0.01	0.59	.03	.07	.05
0.4	1.02	0.30	0.52	2.23	0.02	0.41	-1.11	1.36	0.16	0.10	0.01	0.59	.03	.07	.05
0.6	1.02	0.30	0.52	2.23	0.01	0.60	-1.58	1.83	0.16	0.10	0.01	0.59	-.04	.07	-.01
0.8	1.02	0.30	0.52	2.23	0.02	0.81	-1.84	2.60	0.16	0.10	0.01	0.59	.13	.07	.01
1.0	1.02	0.30	0.52	2.23	0.00	1.00	-2.65	3.11	0.16	0.10	0.01	0.59	.08	.07	-.04
1.2	1.02	0.30	0.52	2.23	0.02	1.21	-2.84	3.30	0.16	0.10	0.01	0.59	-.04	.07	-.08
1.4	1.02	0.30	0.52	2.23	0.01	1.41	-3.62	3.83	0.16	0.10	0.01	0.59	.01	.07	.04
1.6	1.02	0.30	0.52	2.23	0.01	1.59	-4.20	4.86	0.16	0.10	0.01	0.59	.02	.07	.07

Note. The simulated item pool based on the Rasch model had the same distribution of *b*-parameters.

4.2.3 Evaluation criteria

For each condition, the recovery of proficiency estimates and the amount of adaptation were evaluated. For the precision and accuracy of the final proficiency estimates, conditional statistics including bias (CB), standard error of measurement based on test information (CTSEM), and root mean square error (CRMSE) were computed at each proficiency level:

$$CB_j = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{jr} - \theta_j), \tag{4.1}$$

$$CTSEM_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\frac{1}{\sum_{i=1}^{L} I_i(\hat{\theta}_{jr})}}, \tag{4.2}$$

$$CRMSE_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{jr} - \theta_j)^2}\,, \tag{4.3}$$

where $\hat{\theta}_j$ is the final proficiency estimate for the examinee *j*, θ_j is the true proficiency of the examinee *j*, $I_i(\hat{\theta}_j)$ is the Fisher information of the *i*th item at the current estimate, $\hat{\theta}_j$, *L* is the test length, and *R* is the total number of replications (i.e., *R* = 50).

Overall statistics were considered to provide summary information of the recovery aggregated over the proficiency levels. The overall statistics including bias, TSEM, RMSE, and the Pearson correlation between true and final estimates of proficiency (i.e., the fidelity coefficient, $r_{\theta\hat{\theta}}$; McBride, 1977) were computed across all examinees within a single replication, where *N* is the total number of examinees, $\bar{\theta}$ is the mean of true proficiency values over *N* examinees, and $\bar{\hat{\theta}}$ is the mean of final proficiency estimates over *N* examinees:

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{N_r}\sum_{j=1}^{N_r}(\hat{\theta}_{jr} - \theta_j)\right) \tag{4.4}$$

$$TSEM = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{N_r}\sum_{j=1}^{N_r}\frac{1}{\sum_{i=1}^{L} I_i(\hat{\theta}_{jr})}\right)} \tag{4.5}$$

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{N_r}\sum_{j=1}^{N_r}(\hat{\theta}_{jr} - \theta_j)^2\right)} \tag{4.6}$$

$$r_{\theta\hat{\theta}} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\sum_{j=1}^{N}(\theta_j - \bar{\theta})(\hat{\theta}_{jr} - \bar{\hat{\theta}}_r)}{\sqrt{\sum_{j=1}^{N}(\theta_j - \bar{\theta})^2}\sqrt{\sum_{j=1}^{N}\left(\hat{\theta}_{jr} - \bar{\hat{\theta}}_r\right)^2}}\right) \tag{4.7}$$

More importantly, to evaluate the performance of statistical indicators of the amount of adaptation, existing adaptation measures (Kingsbury & Wise, 2018; Reckase et al., 2018) and the conditional adaptation measures I proposed were calculated using the equations listed in (3.2) through (3.8). Furthermore, relationship between the proposed conditional adaption indices and the conditional measurement statistics was visually inspected via a scatter plot.

4.3 Research Question 2

To demonstrate the practical utility of the proposed conditional measures for the amount of adaptation as diagnostic tools for improving the adaptivity of a CAT, this research question investigated how many items need to be added to attain an acceptable level of adaptation over the proficiency continuum under the hypothesized scenario.

Suppose a state-wide achievement testing program is planning to improve an adaptive test with the goal of reaching comparable measurement precision of a student's proficiency

across the entire range of proficiency levels. To do this, the first basic step that they want to take is to revise an item pool that makes a CAT that gives items that match well the examinee's proficiency estimate. According to their history of students' proficiency distributions, the test developers found that students' proficiency generally followed a standard normal distribution, locating many students in the middle proficiency levels. However, given the proficiency population with the currently available item pool, some range of proficiency levels is adequate for the customization of the CAT to students, whereas other areas may not be. To improve adaptivity at the proficiency area that is currently below the criterion, items whose information peaks over that area need to be added to the existing item pool. The items to be added are selected from the master pool, which is usually available in real-world operational settings.

4.3.1 Item pool

The 300-item pool for the 3PL model developed for the item-pool-size study in the first research question was employed as the existing item pool to be revised later. The reason for choosing that pool size is that the item-pool-size study would suggest that a pool size of 300 presented acceptable adaptation, and at the same time, there is still room to approach a better level of adaptation for a fixed-length CAT of 40 items.

Next, the master pool, known as a "vat" (e.g., Way, 1998), was created that has a larger number of items than required by a CAT. I generated 3,000 sets of item parameters for the master pool that mimics the target distribution of item parameters described in Table 4.1, taking into account the correlations among *a*-, *b*-, and *c*-parameters. The resulting distributions of item parameters for the master pool were very similar to the target of Table 4.1. The distributions of *b*-parameters for the master pool is presented in Figure 4.1.

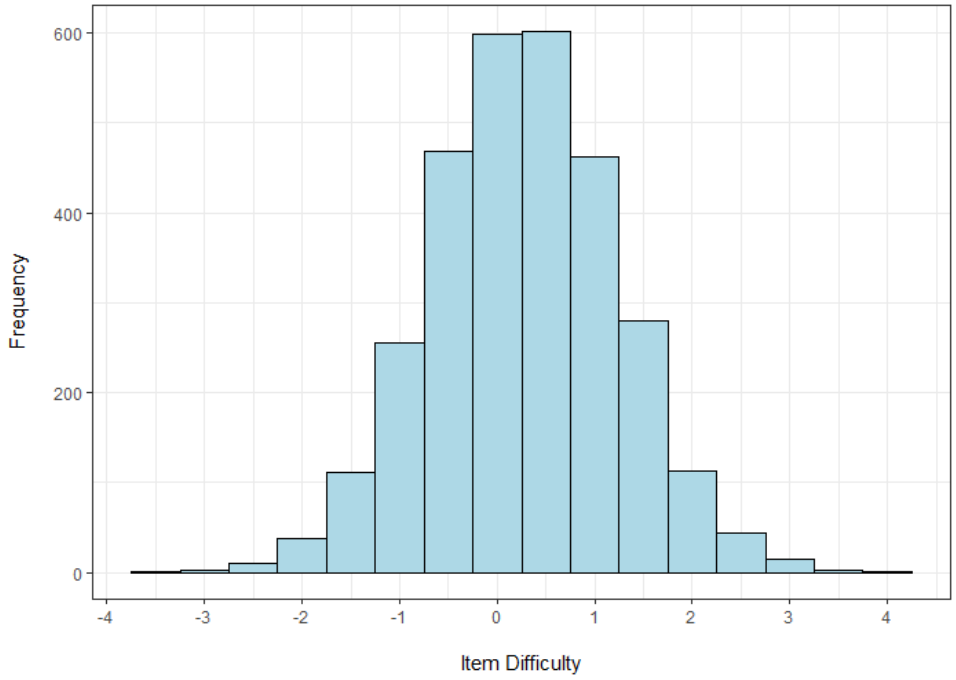


Figure 4.1. Item distribution for the master pool (N = 3,000).

4.3.2 Simulation procedure

First, 2,000 examinees were randomly sampled from a standard normal distribution, *N* (0, 1). A 40-item CAT was then administered to examinees using the 300-item pool with the 3PL model. I then identified the values of the conditional adaptation measures, which were considered as a baseline. Looking at those values over the entire proficiency level range, I determined which region on the proficiency scale needs to improve adaptivity of the CAT giving more appropriate items which are customized to examinees' proficiency level. Then, better targeting items were added that ought to be sufficient to gain the desired level of adaptivity. At each θ-region under the benchmark values of the three statistics, the fixed numbers of items to be added to the item pool are 5, 10, 15, 20, 30, 40, 50, and 100.

4.3.3 Evaluation criteria

To answer the second research question, three conditional measures for the amount of adaptation were computed using equations in (3.6) through (3.8) to see whether the test can be labeled as having "good" adaptation at the area of interest after the items have been added to the existing item pool. Conditional statistics listed in (4.1) to (4.3) were also calculated for checking to what extent the measurement precision of the proficiency estimates was improved after the item pool was revised.

4.4 Research Question 3

Imposing constraints on item selection for exposure control may contribute to reducing the amount of adaptation during a CAT. In Research Question 3, I examined whether the proposed indices can properly capture the changes of adaptivity resulting from the exposure-control procedure over the proficiency continuum.

4.4.1 Simulation design

I designed the effects of exposure-control procedures moderated by item pool designs to emphasize the capability of the indices that identify the differences in adaptivity given by item pool quality and constraints on item selection. In total, there are eight conditions (2 item pools × 4 exposure control procedures). For each condition, CATs were administered to 2,000 examinees randomly sampled from the standard normal distribution, N (0, 1) over 50 replications.

4.4.1.1 Item pool

I created two item pools: (1) an *optimal* operational item pool and (2) a *regular* operational item pool. First, the optimal item pool was designed using the bin-and-union method (see Reckase, 2010 for details on this procedure). The optimal pool has a sufficient number of items with a distribution that satisfies the desired features of a CAT administered to the target

population of examinees (e.g., Veldkamp & van der Linden, 2000). Using the bin-and-union method, the blueprint of the ideal item pool design can be identified in terms of the distribution of items, item characteristics, and item pool size for the predetermined CAT specifications of interest. More specifically, the ideal item pool was first determined by tallying the number of selected items needed in each range, called "bins", of the proficiency estimates, which are specified on the proficiency scale, producing a target distribution for items over bins. The bin size is the median range of near maximum information, which was determined based on having information within 90% of the maximum for an item. In this case, the bin width was 0.7.

To design the ideal pool, through simulations, a 40-item CAT selected from the master pool was administered to 2,000 examinees, and as each CAT is administered and the union of the required items is taken, the ideal item pool grows in size when simulated examinees that are different than those previously selected are chosen. Here, the master pool was the same as one that was already created in Research Question 2. As seen in Figure 4.2, the size of the ideal item pool quickly grows early in this process and then reaches an asymptote once most of students' proficiency range is covered. The number of items at the asymptote is an estimate of the number of items needed for the ideal item pool. Since the simulation is a random process, it was replicated 10 times and then the median pool size and the median value in each of the bins were determined for the ideal item pool. The median of the sizes for the ideal item pool was 400 items.

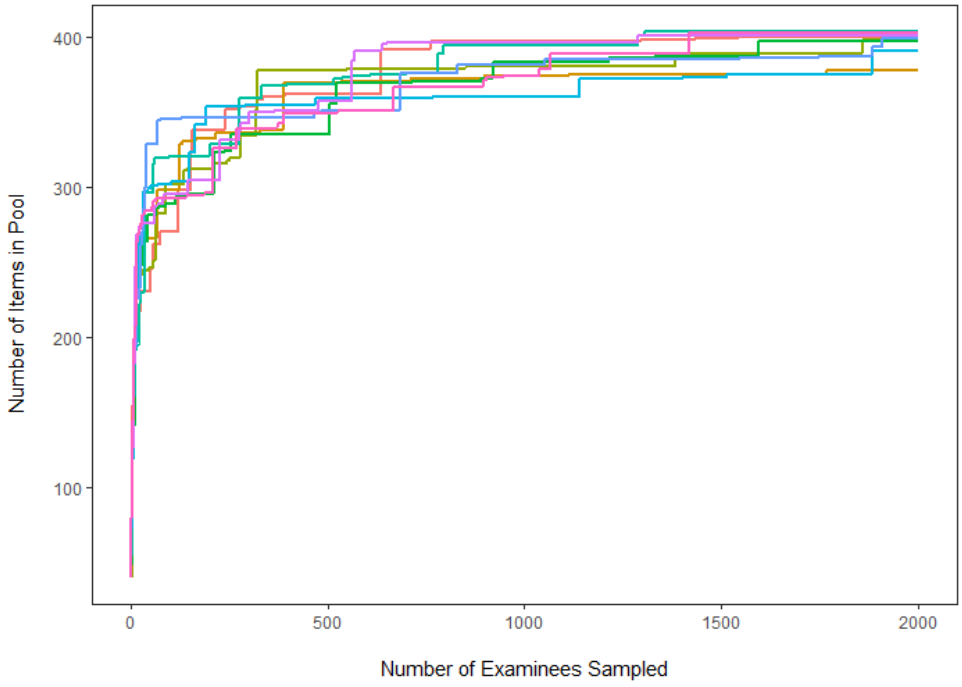


Figure 4.2. Number of items needed in the ideal item pool for a 3PL-based CAT of 40 items.

However, the ideal item pool is sometimes not realistic because it requires items for extremely high or low proficiency levels that are not encountered very often in practice. Therefore, after identifying the distribution of items for the ideal item pool, items from a master pool then filled in the requirements of the frequency distribution over bins in the ideal pool design. Items were selected that had maximum information for the proficiency range defined by the bin boundaries. In some bins, no items were available in the master item pool. The union of the selected items is viewed as the *optimal* operational item pool because this does not contain extremely easy or difficult items, it can be considered reasonably an optimal pool, in practice. The size determined for this optimal pool was 300. To make the two pools of similar size, the regular operational item pool consisted of 300 items. The 300-item pool developed in Research Question 1 was used as a typical operational pool because that pool was generated in a way that mimicked the real item pool from the state-wide testing program.

Table 4.4 presents the distributions of items over bins for the ideal item pool, the optimal item pool, and the regular item pool. Compared to the other two item pools, the ideal item pool had 32 items with maximum information in the -3.85 to -3.15 bin, 35 items with the maximum in the -3.15 to -2.45 bin, and 34 items with the maximum in the 3.15 to 3.85 bin. These items were relatively extreme in difficulty given the distribution of items. Since the master pool did not have sufficient items that had maximum information at such extreme proficiency regions, the optimal operation item pool had 1, 6, and 5 items, respectively in those ranges on the proficiency scale. Other than the extremes, the optimal item pool had almost identical distribution of items to the ideal item pool that included items fairly uniformly distributed from -2.45 to 3.15. Meanwhile, the real item pool had a visibly narrower distribution of items with the largest frequency in the 0.35 to 1.05 range of the proficiency scale. Given the purpose of the test is not to classify students into mastery vs. non-mastery using the single cut-off score but to attain equal measurement precision over the proficiency continuum, at the least the test would need more easy items for the low proficiency students. Despite the same pool size, the optimal pool and the regular pool apparently had a different distribution, which is visualized in Figure 4.3.

(a) Regular Item Pool (b) Optimal Item Pool

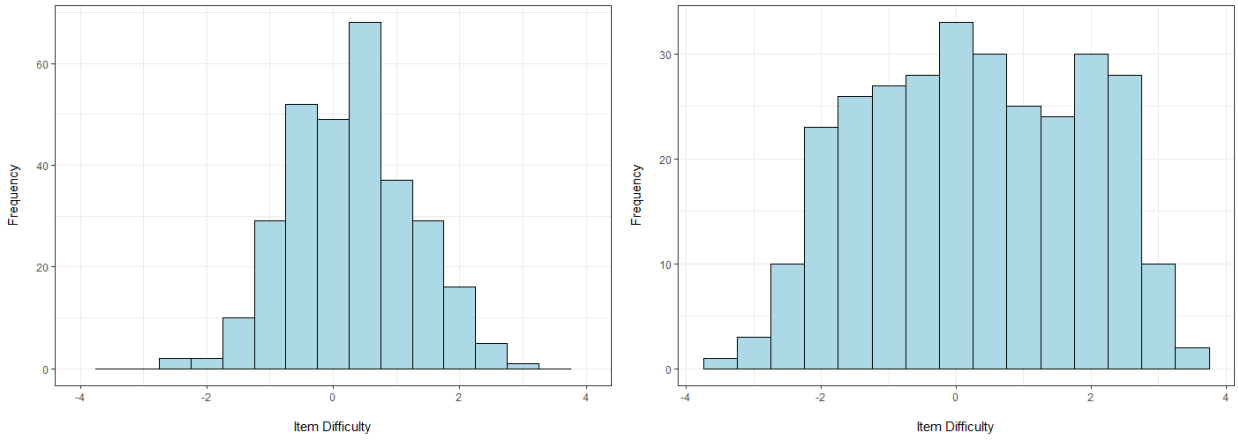


Figure 4.3. Distribution of *b*-parameters for the regular and optimal item pools.

Table 4.4

Item Distributions for Item Pools Considered in Research Question 3

Bin Boundaries (θ-Scale)		Ideal Item Pool	Optimal Item Pool	Regular Item Pool
Lower bound	Upper bound			
-3.85	-3.15	32	1	0
-3.15	-2.45	35	6	0
-2.45	-1.75	37	24	3
-1.75	-1.05	38	38	13
-1.05	-0.35	39	39	53
-0.35	0.35	39	39	73
0.35	1.05	38	38	88
1.05	1.75	38	38	45
1.75	2.45	37	37	23
2.45	3.15	35	35	1
3.15	3.85	34	5	1
	Total	400	300	300

4.4.1.2 Exposure control methods

Along with a no-exposure control as a reference, the study considered three commonly used exposure-control methods. The first exposure-control approach is the randomesque procedure (Kingsbury & Zara, 1989), in which an item to be administered is randomly selected from the *N* items that have the best information at the current proficiency estimate. In this study, one item was selected out of the 10 most informative items at the current proficiency estimate.

The second procedure is the Sympson-Hetter method (Sympson & Hetter, 1985) with a target rate of maximum item exposure, which was 0.20 in this study. This method is a probabilistic item exposure control in CAT by separating the item selection process from the item administration process. Specifically, this approach employs a simulation of the CAT procedure using the actual item pool to determine how often items will be selected for administration given an expected distribution of examinees. In this process, an exposure control

parameter is estimated for each item in the item pool, which is the conditional probability that the item will be administered if that item is selected. The control parameters have to be determined through an iterative process of the CAT until the exposure control parameters are stabilized. In this study, the stable values of the exposure control parameters were obtained after 15 iterations of the CAT process with each of the regular and optimal item pools. Figure 4.4 presents the distribution of the estimated exposure control parameters for the two item pools. For these pools, over 125 items had exposure control parameter values of 1.0, which means that no exposure control was needed for these items. These items might be unused or underexposed in the CAT process. For the regular item pool, over 50 items had the control parameter values of around 0.4, while for the optimal item pool, about 100 items had the control parameters of around 0.3 and 0.4.

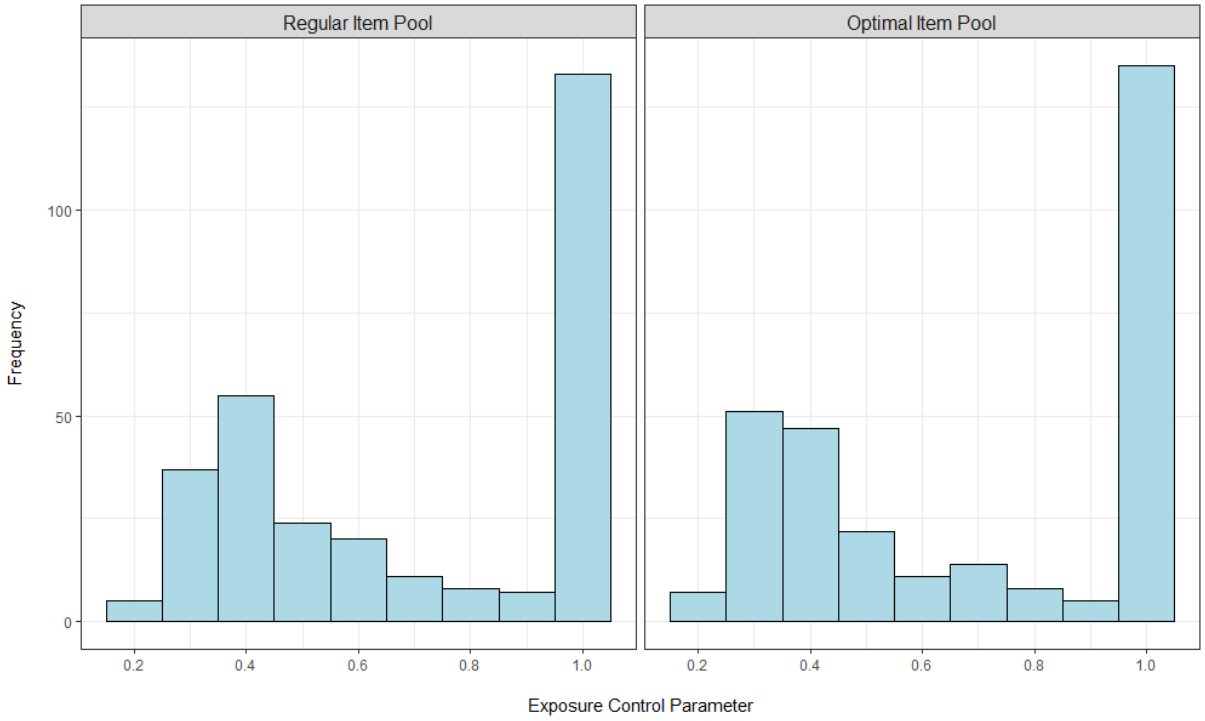


Figure 4.4. Distribution of exposure control parameters for the Sympson-Hetter procedure for the regular item pool (left) and the optimal item pool (right) of 300 items.

Lastly, the *a*-stratified with *b*-blocking procedure (BAS; Chang et al., 2001) was considered for exposure control. For the implementation of BAS, the item pool was partitioned into four levels (strata) based on the magnitude of the *a*-parameters, but the strata were blocked on the *b*-parameter to ensure that the mean and standard deviations for the *b*-parameters were about identical across the four strata. That is, the item pool was first sorted according to the magnitude of the *b*-parameters and divided into 75 groups with each group consisting of four items that were homogeneous in the *b*-parameter. Then, starting with the first block of four items, the item with the lowest *a*-parameter was located in the lowest stratum, the item with the next lowest *a*-parameter in the second stratum, and so on. This procedure continued for each block of items to create four strata of item pools that differed in the magnitude of *a*-parameters but spanned the similar range of *b*-parameters. Note that for BAS, an item was selected with its *b*-parameter closest to the interim estimate of proficiency instead of the MFI item selection.

4.4.2 Evaluation criteria

Similar to previous research questions, conditional statistics in Equations (3.6) through (3.8) and overall statistics for the amount of adaptation in Equations (3.2) through (3.4) were compared across eight conditions along with statistics for evaluating measurement precision and accuracy in the proficiency estimates. In addition, so as to further examine test security, I reported the distribution of observed item exposure rates, computed using Equation (4.8).

$$r_{exposure,i} = \frac{t}{N} \tag{4.8}$$

where *t* is how many times an item *i* was administered and *N* is the total number of examinees.

4.5 Research Question 4

The fourth simulation study investigated the utility of these conditional adaptivity measures to identify the difference in the amount of adaptation that occurs when a MST is used instead of a fully item-level CAT, moderated by different item pool designs. Two study factors were manipulated: (a) item pool design and (b) adaptive test design. Since all factors studied were fully crossed with each other, four conditions (2 item pool × 2 test design) were examined. For each condition, each adaptive test was administered to a simulated sample of 2,000 examinees over 50 replications.

4.5.1 Item pool

As with Research Question 3, two types of item pools were used. One is an *optimal* item pool that had more uniform counts of items across the proficiency levels. The other is a *regular* item pool, which is a bell-shaped distribution of items usually found, in practice. In this study, I used the same item pools that were created in the study for Research Question 3. The regular item pool contained more items whose information peaked in the range of -1.05 to 1.05 on the θ-Scale, whereas the optimal item pool included items of which difficulty were broadly distributed. Again, the size of the two pools was 300.

4.5.2 Test design

The test length for both CAT (i.e., item-level adaptive test design) and 1-2-3 three-stage MST (i.e., module/stage-level adaptive design) was 40 items. A fixed length CAT design with the same specifications as above was employed. Regarding the 1-2-3 MST design (as presented in Figure 4.5), I utilized a single panel with increasing module length, staring with the short module in the routing test and ending with a longer module in Stage 3 (i.e., 10-10-20 design). Prior research (e.g., Kim & Kim, 2018; Reckase et al., 2017; Svetina, Liaw, Rutkowski, &

Rutkowski, 2019) found that administering few items in the beginning stage and more items in the last stage tended to produce more accurate final proficiency estimates.

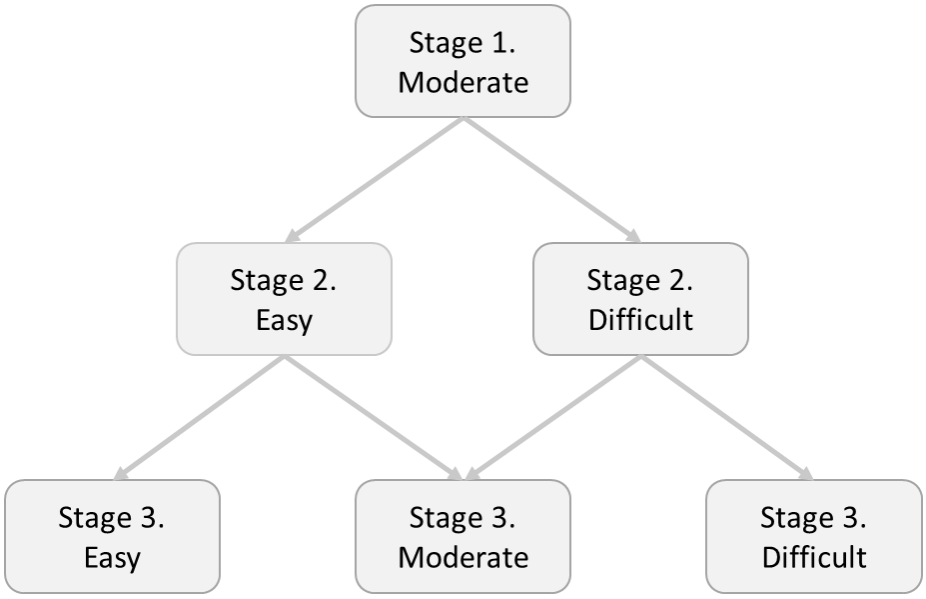


Figure 4.5. A 1-2-3 three-stage MST design used in the study.

For the stage and module configurations, two MST designs were formed from the different item pools. From each of the item pools, a routing module in Stage 1 was constructed so that the test information function (TIF) would match the "target" TIF as closely as possible based on a single decision point of 0.0 to route examinees to one of two second stage modules. The second stage modules were also designed to make accurate classifications of examinees into the three modules in Stage 3 so that items for each second-stage module with TIF peaked at a cut-off point of -1 and 1, respectively, were selected. Lastly, items for the third-stage modules were selected to provide approximately uniform information of the final estimates across the proficiency levels, taking into account the amount of information obtained from the previous stages. Thus, in Stage 3, the easy module was designed for the proficiency (θ) range from -2 to -1, the moderate module was for the θ-range from -1 to 1, and the difficult module was designed for the θ-range from 1 to 2. Table 4.5 displays descriptive statistics of item difficulty parameters

by stage modules for each MST design. The medium module in Stage 3 consisted of relatively easy items, as more informative items were needed to make the information curve flat over the θ-range of -1 to 1.

Among numerous routing strategies, each examinee was routed through modules of which the difficulty levels match the examinee's proficiency level as closely as possible. Examinees were routed based on the IRT MLE proficiency estimate and were not allowed to take non-adjacent paths based on the findings of previous research (Kim & Kim, 2018; Svetina et al., 2019). The TIF function for four possible paths of each MST design in Figure 4.6 showed that the height of the TIF was higher for the MST from the regular pool over the middle range of proficiency (θ), while the breadth of the TIF was broader for the MST formed from the optimal item pool.

Table 4.5

Descriptive Statistics of b-Parameters by Stage for Each MST Design

Stage	Module	Number of Items	*b*-parameters			
			M	*SD*	Min	Max
Regular Item Pool						
Stage 1	Routing	10	-0.02	0.33	-0.62	0.40
Stage 2	Easy	10	-0.93	0.18	-1.21	-0.65
	Difficult	10	0.86	0.27	0.48	1.30
Stage 3	Easy	20	-1.35	0.65	-2.43	-0.15
	Medium	20	-0.29	0.18	-0.60	-0.07
	Difficult	20	1.50	0.14	1.26	1.81
Optimal Item Pool						
Stage 1	Routing	10	0.07	0.22	-0.25	0.33
Stage 2	Easy	10	-0.86	-1.35	-1.18	0.15
	Difficult	10	0.99	0.28	0.62	1.43
Stage 3	Easy	10	-1.66	0.80	-2.58	-0.14
	Medium	10	-0.45	0.33	-1.15	-0.09
	Difficult	20	1.53	0.18	1.19	1.88

Note. Routing points were 0.0 for the first stage module, -1 for the easy module in Stage 2, and +1 for the difficulty module in Stage 2.

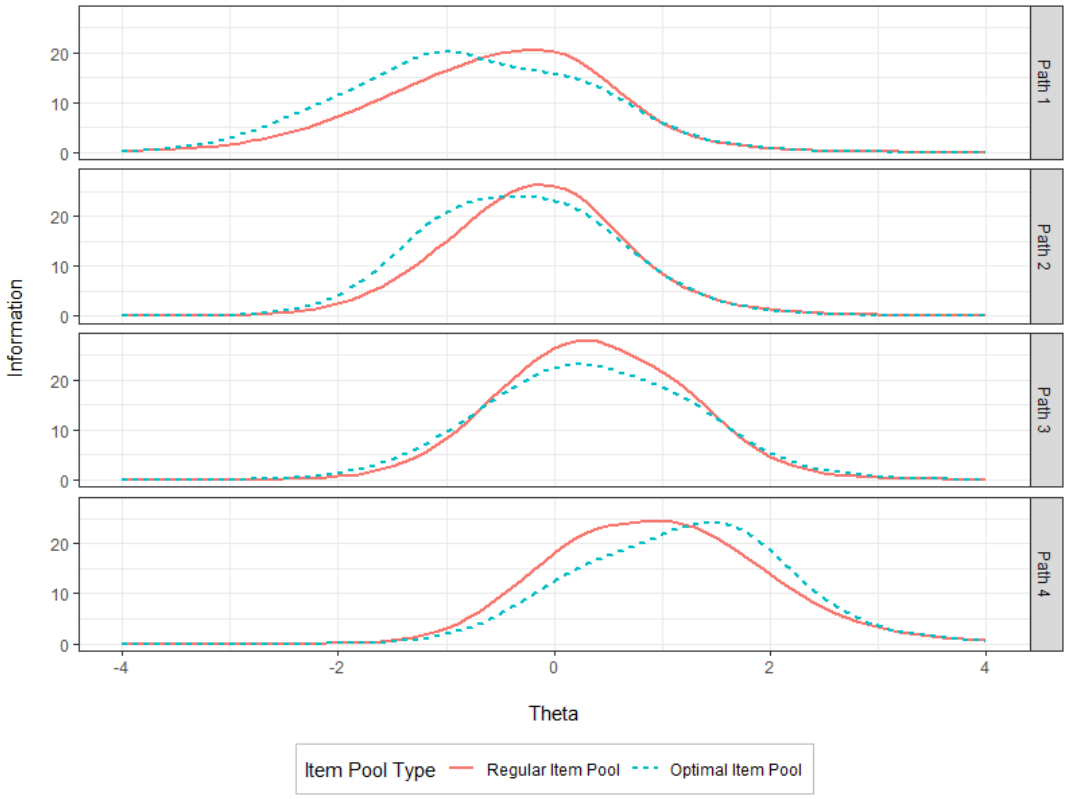


Figure 4.6. Information function by each path for the 10-10-20 MST using regular item pool and optimal item pool.

Note. Path 1 = Stage 1 – Easy in Stage 2 – Easy in Stage 3; Path 2= Stage 1 – Easy in Stage 2 – Medium in Stage 3; Path 3 = Stage 1 – Difficult in Stage 2 – Medium in Stage 3; Path 4= Stage 1 – Difficult in Stage 2 – Difficult in Stage 3.

4.5.3 Evaluation criteria

The performance of the two test designs using different item pools was evaluated in terms of measurement precision and the amount of adaptation. First, I examined how proficiency was accurately and precisely estimated in adaptive testing using bias, TSEM, and RMSE over the sample of examinees, listed in Equations (4.4) to (4.6), and contingent on the proficiency levels in Equations (4.1) to (4.3). The Pearson correlation between "true" and final estimates of proficiencies in Equation (4.7) was also calculated to gauge the relation between true and estimated proficiency. More importantly, the conditional adaptation measures that I proposed in

Equations (3.6) to (3.8) were calculated to investigate the adaptivity at each proficiency level. The existing overall adaptation indices in Equations (3.2) though (3.4) were also computed to understand the adaptivity over the entire group of examinees.

4.6 Research Question 5

In the last question, the performance of the proposed adaptivity measures were examined using real operational CAT data. To do this, the National Council Licensure Examination for Registered Nurse (NCLEX-RN) examination was used. The NCLEX-RN (National Council of State Boards of Nursing [NCSBN], 2017) is a nursing licensure examination delivered by the CAT format, which is administered by NCSBN. This exam assesses "the knowledge, skills, and abilities that are essential for the entry-level nurse to use in order to meet the needs of clients requiring the promotion, maintenance, or restoration of health" (NCSBN, 2016). The full sample for this quarter administration period was about 70,000, which was huge. Instead of computing adaptation statistics using the entire sample, multiple samples of 2,000 examinees were randomly drawn from the total sample, resulting in 35 samples in total. Thus, adaptive measures were computed over 35 samples of 2,000 examinees, allowing for the evaluation of the stability of the adaptation values, as well. In what follows, the details of the NCLEX-RN exam are described in terms of the CAT specifications and the item pool used in this study.

4.6.1 CAT specifications for the NCLEX-RN exam

The NCLEX-RN examination employs the Rasch-based variable-length CAT. On an operational examination, proficiency is estimated using an Owen Bayesian estimation (Vale & Weiss, 1977) with a prior with the mean of -1.0 logit and the standard deviation of 2.0 first until both correct and incorrect responses exist for an examinee. The proficiency estimate is then updated using the MLE with Newton-Raphson. An examinee starts with an item that has 1.0

logit below the cut-off score. The current NCLEX-RN's cut score is 0.0. To pass the examination, the candidate's proficiency estimate should be 1.65 times the standard error (95% confidence interval, one-tail) higher than the cut score. In the same logic, an examinee will fail the exam if their proficiency estimate is 1.65 times the standard error lower than the cut score. Here, the standard error is recalculated after each item. Based on the stopping rule of CAT using a standard error, resulting in a variable-length CAT, the minimum test length of operational items is 60 and the maximum is 250. Each examinee also take additional 15 pretest items in each examination between the 10^{th} and 60^{th} operational items, which are not included in proficiency estimation and analyses in this study. If it cannot be clearly decided whether an examinee passes or fails an examination, the decision then would be made on the basis of the proficiency estimate after taking the final items by examining whether the final proficiency estimate exceeds the cut score of 1.0 to pass.

More importantly, the NCLEX-RN exam has three parts to the content and item selection procedure. First, the computer system determines the number of items for each of eight content strands for the minimum length exam. Every examinee will receive the same number of items per content area for the first 60 items shown in Table 4.6. The second component is that the order of items is determined by randomly selecting a content area with equal probability. Once a content area has been exhausted (e.g., an examinee took the maximum number of items from a category), items from that content area will no longer be tested during the minimum length test. After the minimum length test, the content strand presenting the greatest divergence from the desired testing percentage is selected (Kingsbury & Zara, 1989). The divergence from the desired percentage is computed using the following formula:

$$D = (TPC - \frac{N}{T}) \quad (4.9)$$

where *TPC* is the target percentage for the content strand, *N* is the number of items previously presented from the content area, and *T* is the total number of items previously presented. After determining the content area, 15 items are picked that have maximum information based on examinee's proficiency estimate. Then one out of the 15 items is randomly administered to that examinee.

Table 4.6

Content Distribution of the First 60 Items for the NCLEX-RN in 2016

Content Strand		Number of Items	Target %	Lowest % - Highest %
Content 1	Management of Care	12	20	17-23
Content 2	Safety and Infection Control	7	12	9-15
Content 3	Health Promotion and Maintenance	6	9	6-12
Content 4	Psychosocial Integrity	5	9	6-12
Content 5	Basic Care and Comfort	5	9	6-12
Content 6	Pharmacological and Parenteral Therapies	9	15	12-18
Content 7	Reduction of Risk Potential	7	12	9-15
Content 8	Physiological Adaptation	9	14	11-17
Total		60	100	

4.6.2 Item pool

For this quarter period in 2016, the NCLEX-RN exam used an operational item pool of 1,244 items across eight content areas. Table 4.7 summarizes the descriptive statistics of *b*-parameters for the item pool used in this study. The distribution of *b*-parameters for each content strand was similar across all eight content areas, with the mean of *b*-parameters close to the cut-off score of 0.0. As shown in Figure 4.7, the information for all content areas peaked around 0.0,

indicating that the item pool includes adequately informative items near the cut-score for the NCLEX-RN exam. It was also expected that the amount of information was greater for the content areas consisting of more items.

Table 4.7

Descriptive Statistics of b-Parameters for the NCLEX-RN Item Pool

Content Strand	*b*-parameters				*n*
	M	*SD*	Min.	Max.	
Content 1	0.02	0.85	-2.35	2.22	248
Content 2	0.03	0.83	-2.27	2.19	150
Content 3	0.06	0.83	-2.22	2.30	112
Content 4	0.00	0.84	-2.13	2.23	112
Content 5	0.00	0.79	-2.24	2.07	112
Content 6	0.04	0.81	-2.24	2.17	186
Content 7	0.01	0.84	-2.32	2.22	150
Content 8	0.06	0.83	-2.24	2.19	174
Total	0.03	0.83	-2.35	2.30	1,244

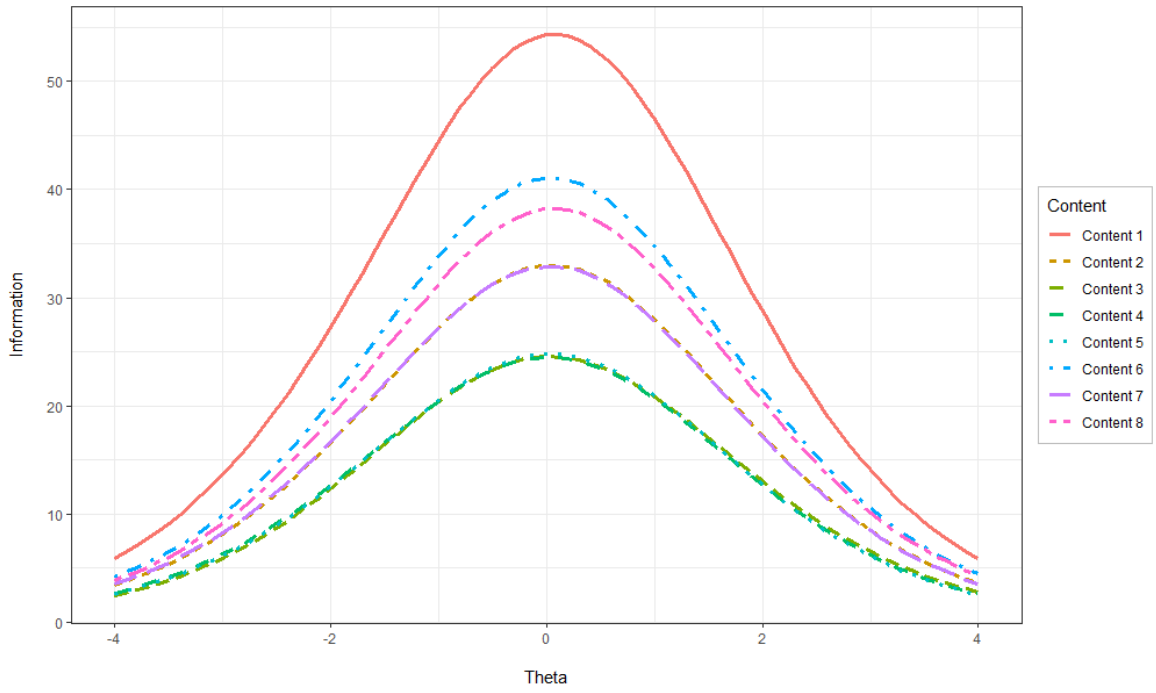


Figure 4.7. Information function by content strand for the NCLEX-RN item pool.

4.6.3 Evaluation criteria

To investigate the amount of adaptation for the NCLEX-RN exam during the quarter of a year in 2016 (April to June), three conditional adaptivity indices listed in Equations (3.6) to (3.8) and three overall adaptivity indices were computed to evaluate adaptivity at individual proficiency level and over the entire sample of examinees.

CHAPTER 5.

RESULTS

This chapter summarizes the results of the analyses organized into five sections corresponding to the five research questions described in Chapter 1. The first four sections present the results of the comprehensive simulation studies that investigated the feasibility and utility of the proposed adaptivity indices conditional on examinees' proficiency levels under numerous conditions. The last section illustrates the empirical demonstration for the conditional adaptivity indices using the real data from the licensure and certification exam.

5.1 Research Question 1

In the first study, comprehensive simulations were conducted to examine the sensitivity of the proposed adaptivity statistics to item pool characteristics, IRT models, and proficiency estimation methods. In particular, item pool characteristics related to the quality of an item pool, were manipulated by (1) varying item pool size and (2) varying item pool spread. These manipulations help understand how well CATs use the available item pool to match item characteristics to each examinee's location on the proficiency continuum. The following two major sections summarize the impacts of these two aspects across different IRT models and proficiency estimation methods in terms of the amount of adaptation as well as measurement accuracy and precision. For each of the studied conditions, all results of the 40-item CATs were averaged across 50 replications.

5.1.1 Variation in item pool size

First, item pool size was investigated to see whether three new conditional adaptation statistics sensitively identify the differences in the amount of adaptation for the CATs using 10

item pools that differed in size from 50 to 500 in increments of 50. This section describes the effect of item pool size across two proficiency estimators (MLE and EAP) when each of IRT models (Rasch model and 3PL model) was used in the CATs, respectively. To better understand the new indices, measurement accuracy and precision were first inspected, followed by the amount of adaptation.

5.1.1.1 Rasch model

Measurement accuracy and precision. Examinees' final proficiency estimates were evaluated using conditional and overall statistics for measurement accuracy and precision. The smaller bias, test-information-based standard error of measurement (TSEM), and root mean square error (RMSE) are, the better the recovery of proficiency estimates is. Also, higher correlation between true and final proficiency estimates ($r_{\theta\hat{\theta}}$) is associated with better recovery.

Conditional statistics. Figure 5.1 shows the mean bias, TSEM, and RMSE across evenly-spaced bins on the proficiency (θ) continuum. The MLE approach presented little bias in proficiency estimates, while the EAP approach reported bias regressing the proficiency estimates toward the mean of prior distribution over the entire θ-continuum. With the EAP, in other words, the proficiency estimates were underestimated at the positive extremes of the θ-scale but overestimated at the negative extremes. The TSEM and RMSE values displayed the slight U-shape pattern showing the higher standard errors at the extremes of θ-continuum than at the moderate proficiency region. The degree of the U-shape pattern was obviously greater for the EAP compared to the MLE. Furthermore, with the bigger item pool, the proficiency estimates appeared to be better recovered, especially for the extreme ends of the scale regardless of their estimation approaches.

Overall statistics. Table 5.1 presents the summary information about overall accuracy and precision of proficiency estimates over the entire sample of examinees. As the pool size increases, the values for bias, TSEM, and RMSE were small, and the correlation coefficients were high. Although the correlations were almost identical between MLE and EAP, EAP produced slightly higher overall bias but lower overall standard errors compared to MLE across the 10 item pools. The differences were getting negligible with larger item pools, though.

Table 5.1

Overall Statistics of Measurement Precision of Proficiency Estimates for a Rasch-based CAT by Item Pool Size and Proficiency Estimator

Pool	MLE				EAP			
Size	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
50	-0.002	0.366	0.371	0.941	-0.006	0.356	0.338	0.942
100	-0.004	0.338	0.342	0.948	-0.002	0.331	0.317	0.949
150	0.000	0.332	0.336	0.949	-0.004	0.327	0.314	0.950
200	-0.001	0.330	0.333	0.950	-0.001	0.325	0.313	0.951
250	-0.001	0.329	0.332	0.950	-0.003	0.325	0.312	0.951
300	0.000	0.328	0.332	0.950	-0.003	0.324	0.311	0.951
350	-0.001	0.328	0.331	0.950	-0.002	0.324	0.311	0.951
400	0.001	0.328	0.330	0.950	-0.003	0.324	0.312	0.951
450	0.000	0.327	0.329	0.951	-0.002	0.323	0.310	0.951
500	-0.003	0.327	0.329	0.951	-0.001	0.323	0.311	0.951

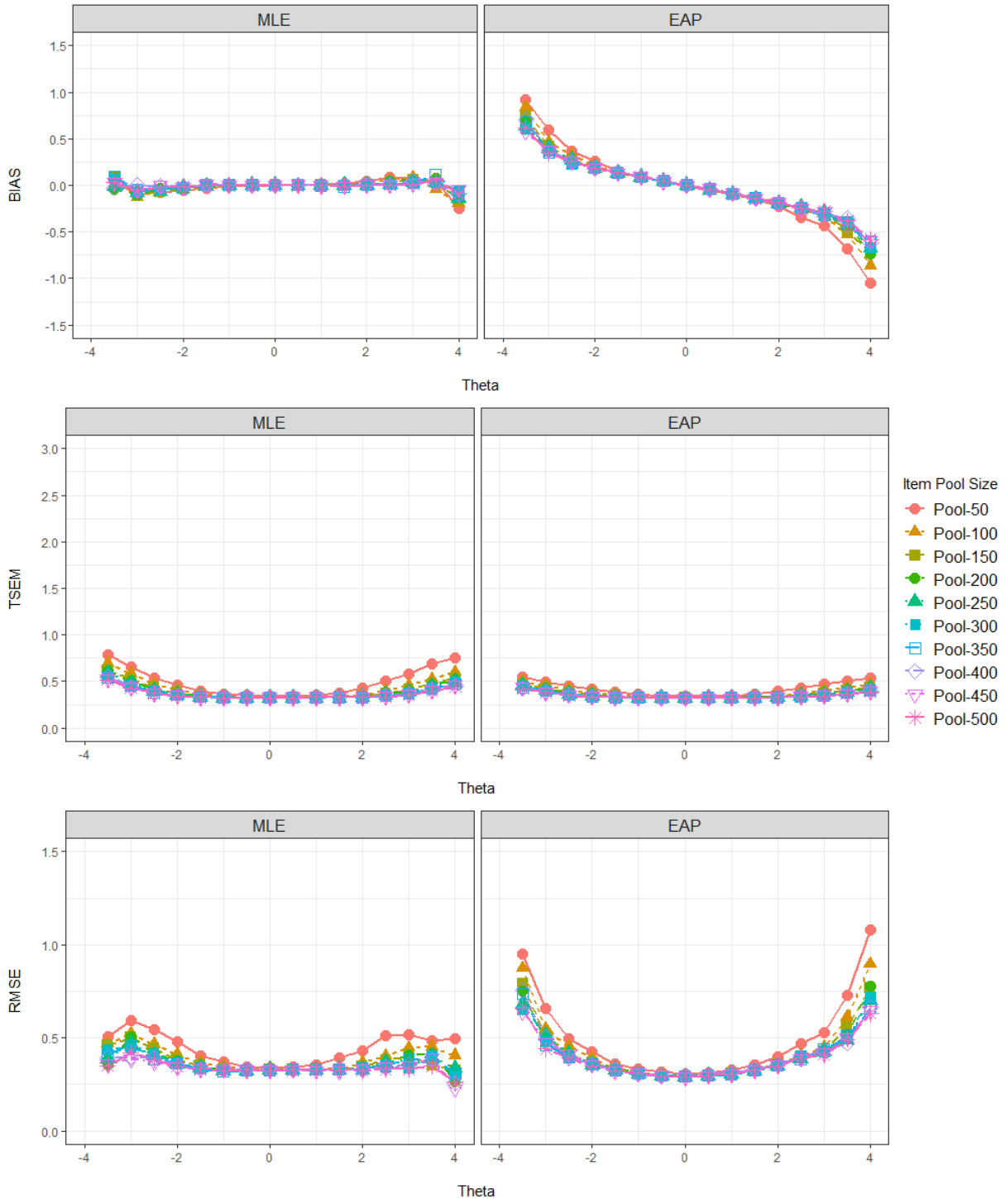


Figure 5.1. Conditional bias, TSEM, and RMSE of proficiency estimates for a Rasch-based CAT by item pool size and proficiency estimator.

Amount of adaptation. Adaptivity for CATs were evaluated over the entire sample of examinees using the existing overall statistics, as well as at the individual proficiency levels using the conditional statistical indicators proposed in this dissertation.

Conditional adaptivity. To evaluate the amount of adaptation for CAT contingent on the proficiency levels, three adaptation statistics were proposed in this dissertation, including deviation of difficulty (DOD), conditional proportion of reduction in variance (CPRV), and ratio of information (ROI) indices. As shown in Figure 5.2, all three adaptivity measures did sensitively detect differences in the amount of customization across the proficiency levels for the CATs using the varying sizes of the item pools. That is, the increase in the item pool size led to higher values of the DOD, CPRV, and ROI indices, implying better adaptation. However, for a given 40-item CAT and its administration procedure, there was not much improvement in adaptivity for pool sizes greater than 300. The POI index (Kingsbury & Wise, 2018) exhibited little sensitivity to item pool sizes and proficiency estimation methods (see Figure 5.2). Additionally, based on the ribbon representing the empirical standard error in Figure 5.2, the values of three adaptation indices appeared to be stable over the 50 replications, though the CPRV index was relatively less precise.

Focusing on the individual proficiency levels, moderate proficiency ranging from -0.5 to 1.5 on the θ-scale produced better adaptivity compared to other proficiency levels across the 10 item pools. This implies that the CAT appropriately selected relatively good items adapted to the students' proficiency estimates from an item pool, and also the item pool contained enough informative items for the students whose proficiency were around the middle level. Interestingly, using the smallest item pool of 50 items reported a slightly different pattern of the three measures compared to other item pools. Although the ROI and DOD indices showed approximately a

symmetric pattern, the CPRV values were asymmetrically distributed. For instance, Student A ($\theta = -1$) took the items whose difficulty varied more than Student B ($\theta = 1$) did, implying that the item pool contained relatively more informative items for Student B compared to those for Student A.

Regarding the comparison of results between the MLE and EAP approaches, the DOD and ROI measures looked more sensitive to the MLE, as the range of their values over the proficiency continuum was greater for the MLE. Also, the values of these two indices for the EAP were generally larger than those for the MLE especially at the extreme regions of the proficiency scale, which is associated with the features of the EAP estimator in terms of the accuracy of proficiency estimates as well as the distribution of the item pool. That is, EAP proficiency estimates were more under- or overestimated at the extreme regions (see Figure 5.3), and the given bell-shaped item pool included fewer good items for the very high or low proficiency levels. Accordingly, their biased proficiency estimates provided more chances of administering informative items to students during the CAT administration. This is evidence that the adaptivity of CAT is closely related to the performance of the proficiency estimation. Meanwhile, CPRV presented a similar pattern between the two estimators, but MLE reported generally lower values with greater empirical standard errors compared to EAP. The latter can be also explained by the property of the proficiency estimation. The MLE used a step size of 0.7 until the examinee had both correct and incorrect responses in the beginning of the CAT, resulting in selecting more heterogeneous items for determining the approximate location of proficiency earlier in the test. The EAP, however, had the benefit of using a prior for the estimation earlier in the test but the biased estimates regressed toward the mean of the prior

distribution, leading to CPRV values that were more stable over the replications and higher in the broader proficiency regions.

In summary, these results suggest that a value in the mid 0.90s for the DOD index, a value in the high 0.70s for the CPRV, and a value in the high 0.90s for the ROI index indicate good adaptation when the Rasch model is used for the CATs with MLE. For the CATs using the Rasch model and EAP, a value in the high 0.90s for the DOD, and a value in high .80s for the CPRV, and a value in the high 0.90s indicate good adaptation.

Overall adaptivity. Table 5.2 reported the results of overall adaptation measures. As expected, the increase in the item pool size was more likely due to increasing the values of the correlation, ratio of SDs, and PRV indices. Similar to conditional adaptivity indices, adaptivity was not much improved for item pool sizes larger than 300 items. However, the POI index yielded an unexpected pattern; as the pool size increases, the POI index decreases. It may be due to the fact that with the smaller item pool, there may be little variation in the administered test items selected between using the provisional proficiency estimate and the final/true proficiency estimate, leading to the POI value of 1.0. Regarding the proficiency estimation methods, the ratio of SDs index was smaller for EAP than for MLE due to the property of the EAP estimates shrinking to the mean of a prior distribution, allowing the CAT to select relatively more homogeneous items for each examinee. On the contrary, other indices including the correlation, PRV, and POI measures were slightly higher for EAP.

MLE results for the 40-item test provided evidence to support the benchmark values from a previous study with the 30-item test (Reckase et al., 2018): low 0.90s for the correlation index, mid 0.80s for the ratio of SDs index, and about 0.80 for the PRV index. EAP results for the 40-item test additionally suggest some benchmark values for interpreting these overall indices for

CATs; a value in the mid 0.90s for the correlation index, a value in the high 0.70s for the ratio of SDs index, and a value in the high 0.80s for the PRV can be considered good adaptation.

Relationship between conditional adaptivity and measurement precision. To visualize their relationships, the TSEM values were plotted against the DOD, CPRV, and ROI measures, respectively by item pool size and proficiency estimator (see Figure 5.4). Within each item pool size condition, the DOD and the ROI measures showed similarly a negative and curvilinear association with the standard errors (TSEM). Despite their nonlinearity, the Pearson correlation coefficients were computed for information purposes only, yielding greater than .90 across all conditions. However, the CPRV measure did not display an obvious linear or curvilinear relation with the standard errors. Instead, there were two lines identified for MLE and EAP, which may be due to the larger TSEM values at the positive and negative extremes of proficiency (remember, the *U*-shaped distribution of TSEM) but relatively constant CPRV values at these regions. A clear finding is that the standard errors were apparently small when CPRV is greater than or equal to its benchmark value, or vice versa. In addition, plots for the relation of RMSE and the adaptation indices are provided in Appendix A. Because RMSE considers both bias and standard errors, their relationships were not as apparent as those with TSEM.

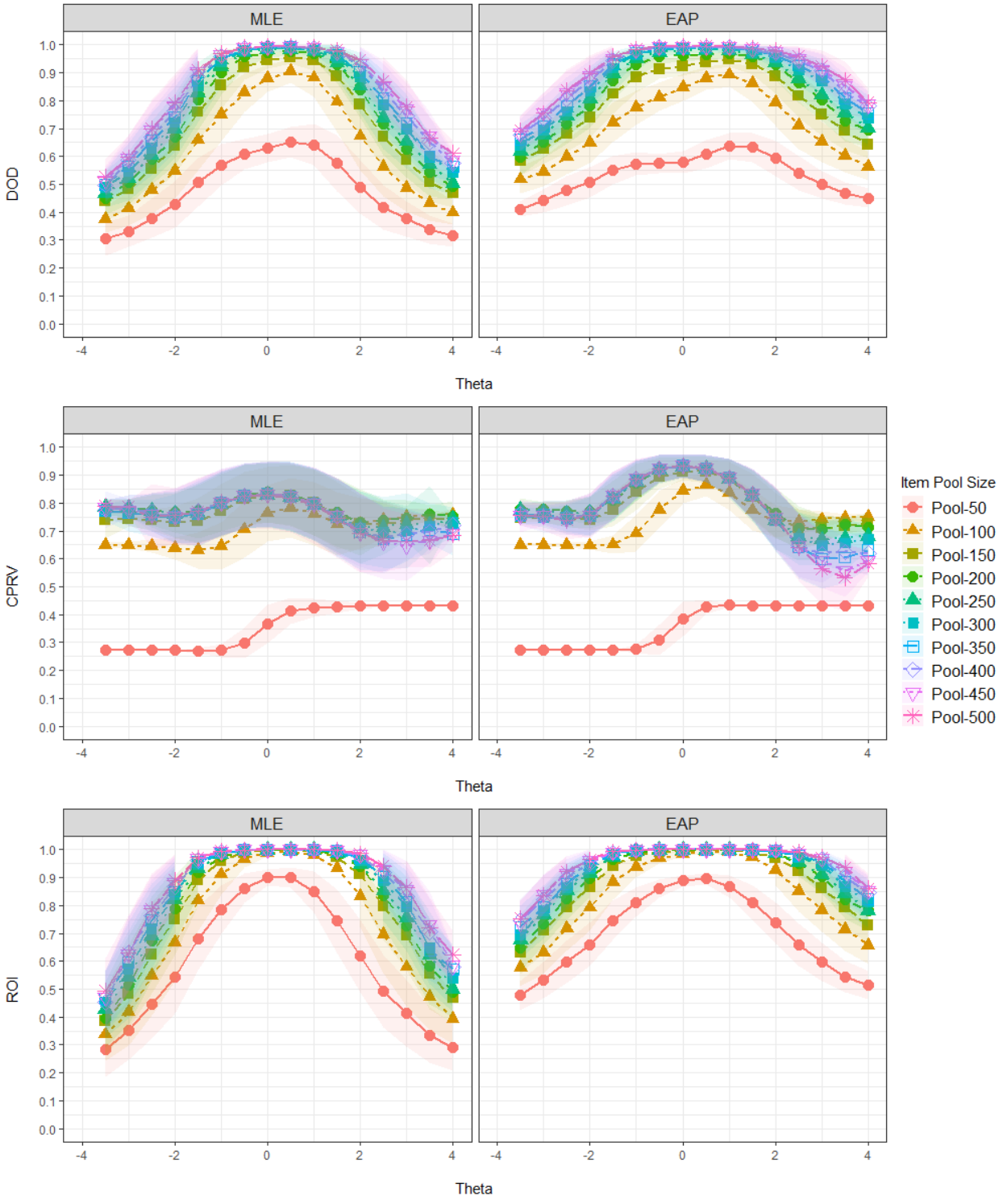


Figure 5.2. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a Rasch-based CAT by item pool size and proficiency estimator.

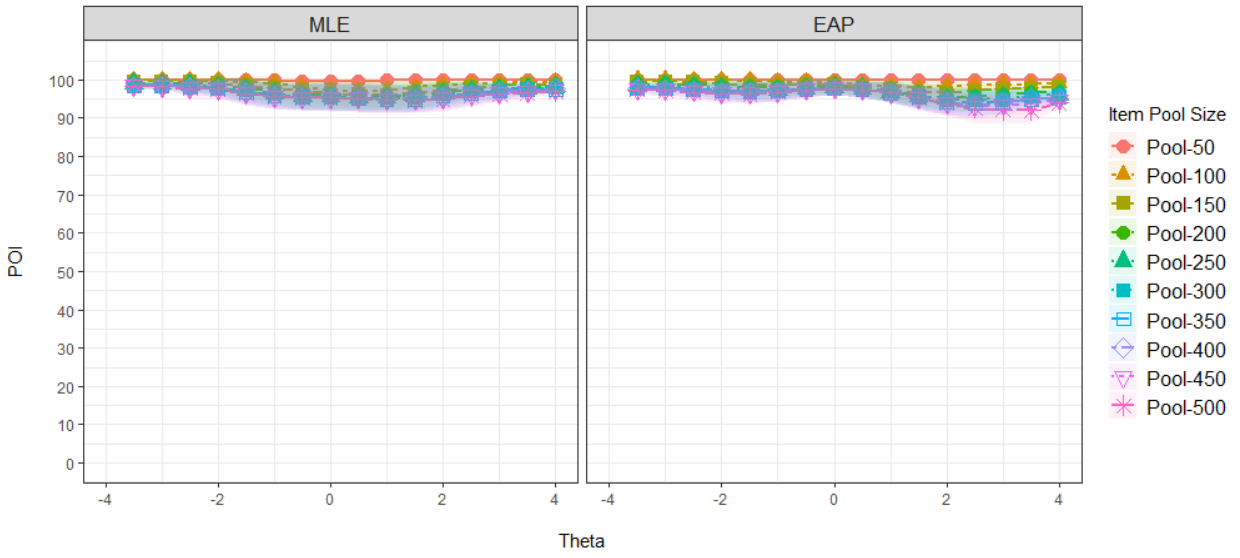


Figure 5.3. Plot of a POI index for a Rasch-based CAT by item pool size and proficiency estimator.

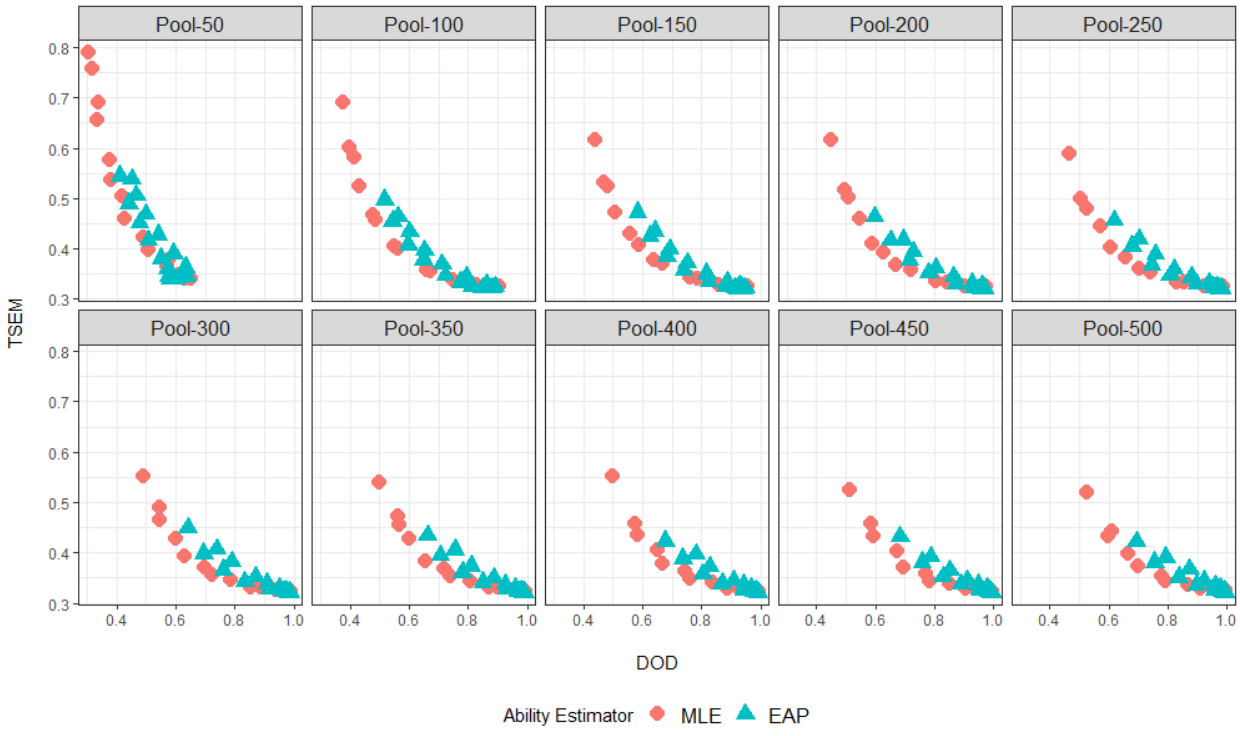


Figure 5.4. Relationship of TSEM with conditional adaptivity indices (DOD, CPRV, and ROI) for a Rasch-based CAT by item pool size and proficiency estimator.

Figure 5.4. (cont'd)

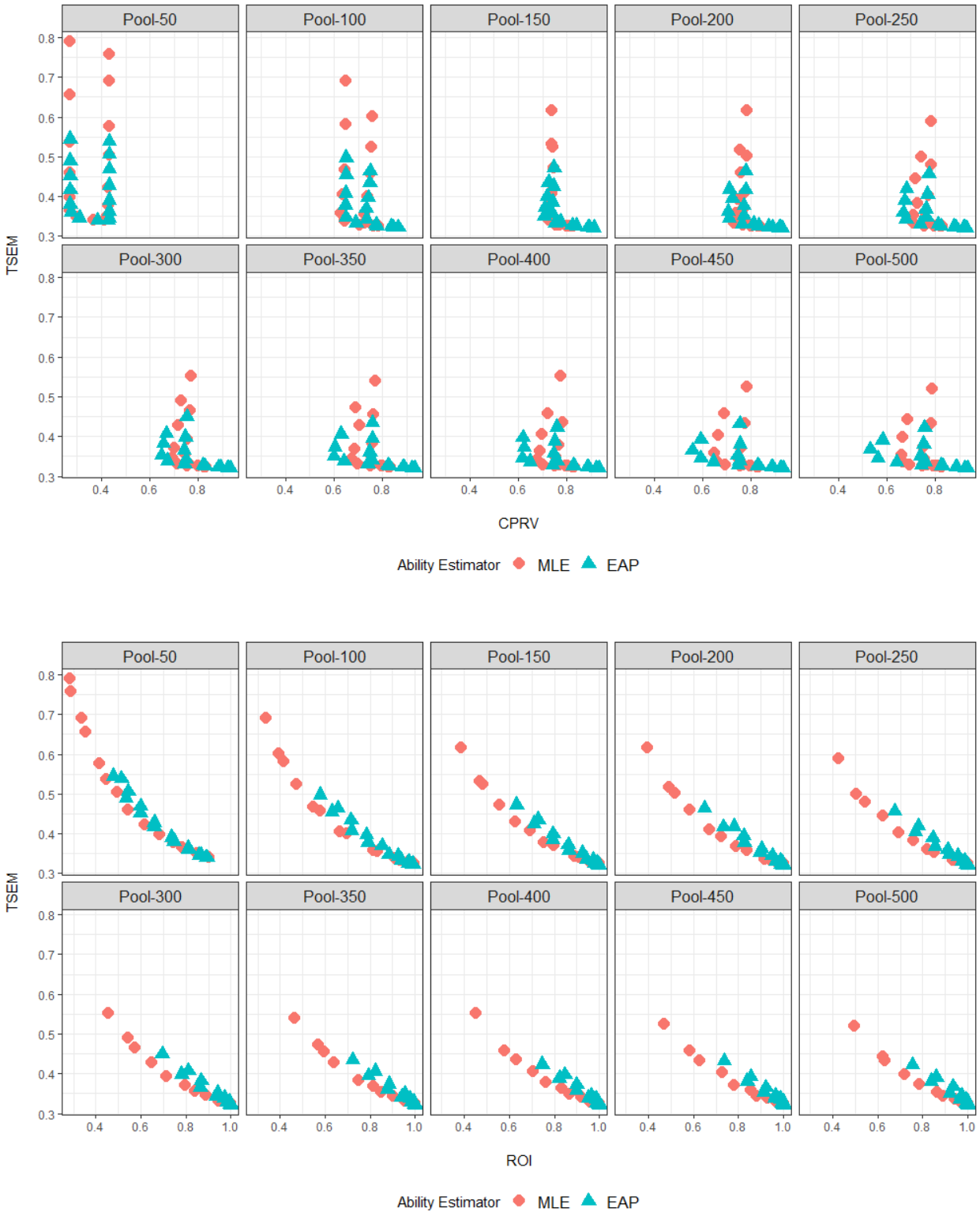


Table 5.2

Overall Adaptation Statistics for a Rasch-based CAT by Item Pool Size and Proficiency Estimator

Pool Size	MLE				EAP			
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI
	M (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
50	0.86 (.003)	0.27 (.003)	0.36 (.001)	99.79 (.016)	0.89 (.003)	0.31 (.003)	0.37 (.001)	99.97 (.004)
100	0.90 (.003)	0.61 (.006)	0.73 (.002)	97.91 (.060)	0.95 (.002)	0.62 (.004)	0.79 (.001)	99.03 (.033)
150	0.92 (.002)	0.74 (.006)	0.80 (.002)	96.88 (.063)	0.96 (.002)	0.71 (.005)	0.87 (.001)	98.25 (.049)
200	0.93 (.002)	0.78 (.006)	0.81 (.003)	96.32 (.075)	0.96 (.001)	0.74 (.005)	0.89 (.001)	97.70 (.052)
250	0.93 (.002)	0.81 (.005)	0.80 (.003)	95.99 (.071)	0.96 (.001)	0.76 (.005)	0.88 (.001)	97.44 (.049)
300	0.94 (.002)	0.83 (.007)	0.80 (.003)	95.77 (.079)	0.96 (.001)	0.78 (.004)	0.89 (.001)	97.24 (.051)
350	0.94 (.002)	0.85 (.007)	0.80 (.003)	95.60 (.065)	0.96 (.001)	0.78 (.005)	0.89 (.002)	97.11 (.052)
400	0.94 (.002)	0.86 (.007)	0.80 (.003)	95.51 (.085)	0.96 (.001)	0.79 (.005)	0.89 (.002)	97.01 (.051)
450	0.94 (.002)	0.88 (.005)	0.80 (.003)	95.41 (.070)	0.96 (.001)	0.80 (.005)	0.89 (.001)	96.95 (.054)
500	0.94 (.002)	0.88 (.006)	0.80 (.003)	95.35 (.086)	0.96 (.001)	0.80 (.004)	0.89 (.001)	96.89 (.041)

Note. All statistics were computed using final estimated proficiencies (θ).

5.1.1.2 3PL model

Measurement accuracy and precision. Like the results for the Rasch model, examinees' final proficiency estimates were evaluated using conditional and overall statistics.

Conditional statistics. Figure 5.5 shows the mean bias, TSEM, and RMSE across evenly-spaced bins on the proficiency (θ) continuum. Unlike the findings for the Rasch model, both the MLE and EAP approaches presented more bias and large standard errors of proficiency estimates at the extreme ends of the proficiency continuum due to the adverse effect of *c*-parameters on the estimations (Thissen & Wainer, 1982). Specifically, the EAP approach for the 3PL model reported greater regress-toward-the-mean bias, whereas the MLE yielded much greater standard errors (TSEM) at the proficiency extremes. While there was more bias of proficiency estimates, implying less accurate estimates for the EAP than for the MLE model, there were large standard errors (TSEM) in the estimates, implying less precision for MLE than for EAP. Considering both bias and standard errors, the RMSE values were higher for EAP at the extreme proficiency regions but they were similar to each other at the middle ranges of the proficiency scale. Moreover, with the smaller item pool, the proficiency estimates appeared to be less accurate and less stable regardless of their estimation approaches.

Overall statistics. Overall accuracy and precision of proficiency estimates were summarized in Table 5.2. As the pool size increased, the values for bias, TSEM, and RMSE decreased, and the correlation coefficients increased. Although the correlations and mean bias were similar to one another, the MLE yielded higher standard errors (TSEM), which results in larger RMSE values across the 10 item pools compared to the EAP.

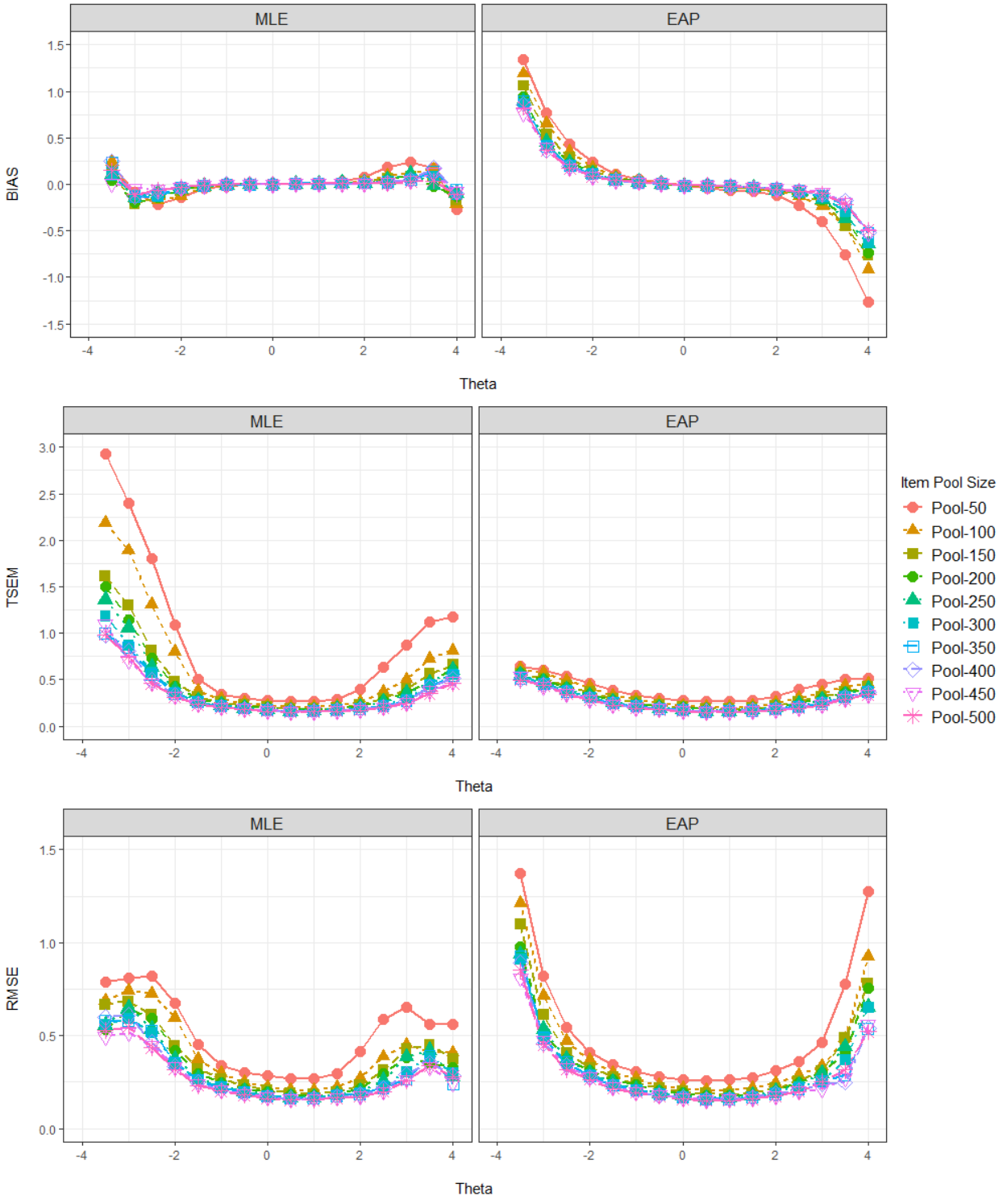


Figure 5.5. Conditional bias, TSEM, and RMSE of proficiency estimates for a 3PL-based CAT by item pool size and proficiency estimator.

Table 5.3

Overall Statistics of Measurement Precision of Proficiency Estimates for a 3PL-based CAT by Item Pool Size and Proficiency Estimator

Pool Size	MLE				EAP			
	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
50	-0.003	0.429	0.345	0.950	-0.003	0.307	0.295	0.956
100	-0.007	0.324	0.280	0.966	-0.002	0.253	0.245	0.970
150	-0.005	0.255	0.245	0.973	-0.002	0.228	0.224	0.975
200	-0.002	0.237	0.231	0.976	-0.001	0.216	0.212	0.978
250	-0.002	0.219	0.213	0.979	-0.001	0.202	0.199	0.980
300	-0.003	0.209	0.207	0.980	-0.002	0.197	0.194	0.981
350	-0.001	0.203	0.203	0.981	-0.001	0.192	0.189	0.982
400	-0.001	0.196	0.196	0.982	-0.001	0.186	0.185	0.983
450	0.000	0.193	0.193	0.982	-0.001	0.184	0.183	0.983
500	-0.001	0.187	0.188	0.983	-0.001	0.179	0.178	0.984

Amount of adaptation. Adaptivity for CATs were evaluated using the existing overall statistics as well as conditional statistical indicators proposed in this dissertation.

Conditional adaptivity. Results indicated that the three measures appeared to be sensitive to variation in item pool size across the proficiency levels shown in Figure 5.6. As the pool size increased, all three measures showed higher values, indicating better adaptation. For the given 40-item test and CAT administration procedure, there was not much improvement in three adaptivity indices for pool sizes greater than 300. Note that, compared to the patterns for the Rasch model, all the values of three adaptation measures were smaller across the proficiency levels. The middle proficiency area produced better adaptivity than other proficiency areas. Not only that, but with the smallest item pool of 50 items, the ROI and DOD indices showed approximately a symmetric pattern centered on 0.0, whereas the CPRV values were asymmetrically distributed.

In addition, according to the shading ribbon representing the empirical standard error of the adaptation measures in Figure 5.6, the values of three adaptation indices appeared to be stable over the 50 replications, but there were relatively higher empirical standard errors in the measures at the very ends of the proficiency continuum. The latter might be due to the limited items available for the extreme or due to the larger standard errors of the proficiency estimates at the very high or low proficiency levels, which is more likely affected by *c*-parameters. However, the POI index was neither sensitive to variation in item pool size nor to proficiency estimators in Figure 5.7.

In comparison to the EAP approach, three adaptation measures appeared to be more sensitive to the MLE across the 10 item pools, and the empirical standard errors of these measures were larger for MLE. These might be related to the measurement properties of the abilities estimators' capability to handle the unstable estimation issues caused by the *c*-parameters when the 3PL model is used. In particular, the CPRV presented a similar pattern between the two estimators, but MLE reported slightly lower values with the greater empirical standard errors compared to EAP. The latter can also be explained by the properties of the proficiency estimation. Along with the *c*-parameter issue, MLE used a step size of 0.7 until MLE can be computed earlier in the CAT, resulting in selecting more heterogeneous items for determining the approximate location of the proficiency level. EAP, however, had the benefit of using a prior for the estimation earlier in the test but presented biased estimates regressed toward the mean of the prior distribution, leading to the CPRV values that were more stable over the replications and higher in the broader proficiency regions.

Taken all together, these results suggest some guidelines for interpreting the adaptation measures; a value in the mid 0.70's is good for DOD, a value in the low 0.80's is considered a

good adaptation for the CPRV, and a value in the mid 0.80's is good for ROI when the 3PL model is used for scaling and scoring with MLE. Results of the 3PL CATs using EAP support the guidelines found using Rasch model, but they suggest a slightly higher benchmark value for the CPRV, which is a value in the mid 0.80's.

Overall adaptivity. Table 5.4 presents the findings of the overall adaptation measures. As item pool size increased, all three measures increased, but these values were not much improved for item pool sizes larger than 300. However, as identified in the pattern for the Rasch model, the POI index decreased, as item pool size increased. Unlike the Rasch model, the differences in the overall adaptivity measures between MLE and EAP were relatively small. While EAP presented slightly higher PRV values across the pool size conditions, other overall measures performed similarly regardless of the proficiency estimators, which is consistent with the findings from a previous study (Ju & Lee, 2018). Furthermore, these results confirmed again the benchmark values for interpreting the overall adaptation statistics suggested from previous studies (Ju & Lee, 2018; Kim et al., 2018). A value in the high 0.90s for the correlation index and a value in the high 0.70s for the ratio of SDs index can be considered good adaptation for the CAT regardless of the proficiency estimation methods. However, the benchmark value for the PRV index is a value in the low 0.80s for the CATs with combination of 3PL/MLE and a value in the mid 0.80s for 3PL/EAP. Again, as with the CPRV index, which is a modified version of PRV, the PRV index might be affected by the properties of proficiency estimation approaches.

Relationship between conditional adaptivity and measurement precision. For brevity, plots of relations of the standard errors (TSEM) with the DOD, CPRV, and ROI indices are displayed in Figure 5.8, which were similar to the relations found using the Rasch model. For each item pool, the DOD and ROI measures were negatively related with the standard errors,

showing a slightly nonlinear pattern for both MLE and EAP. In spite of their nonlinearity, the Pearson correlation coefficients were computed for information purposes only, reporting very strong, negative relationships between the TSEM and either DOD or ROI. However, again, the CPRV measure did not present an apparent pattern for the relation of standard errors. The relations between RMSE and adaptivity indices are presented in Appendix A.

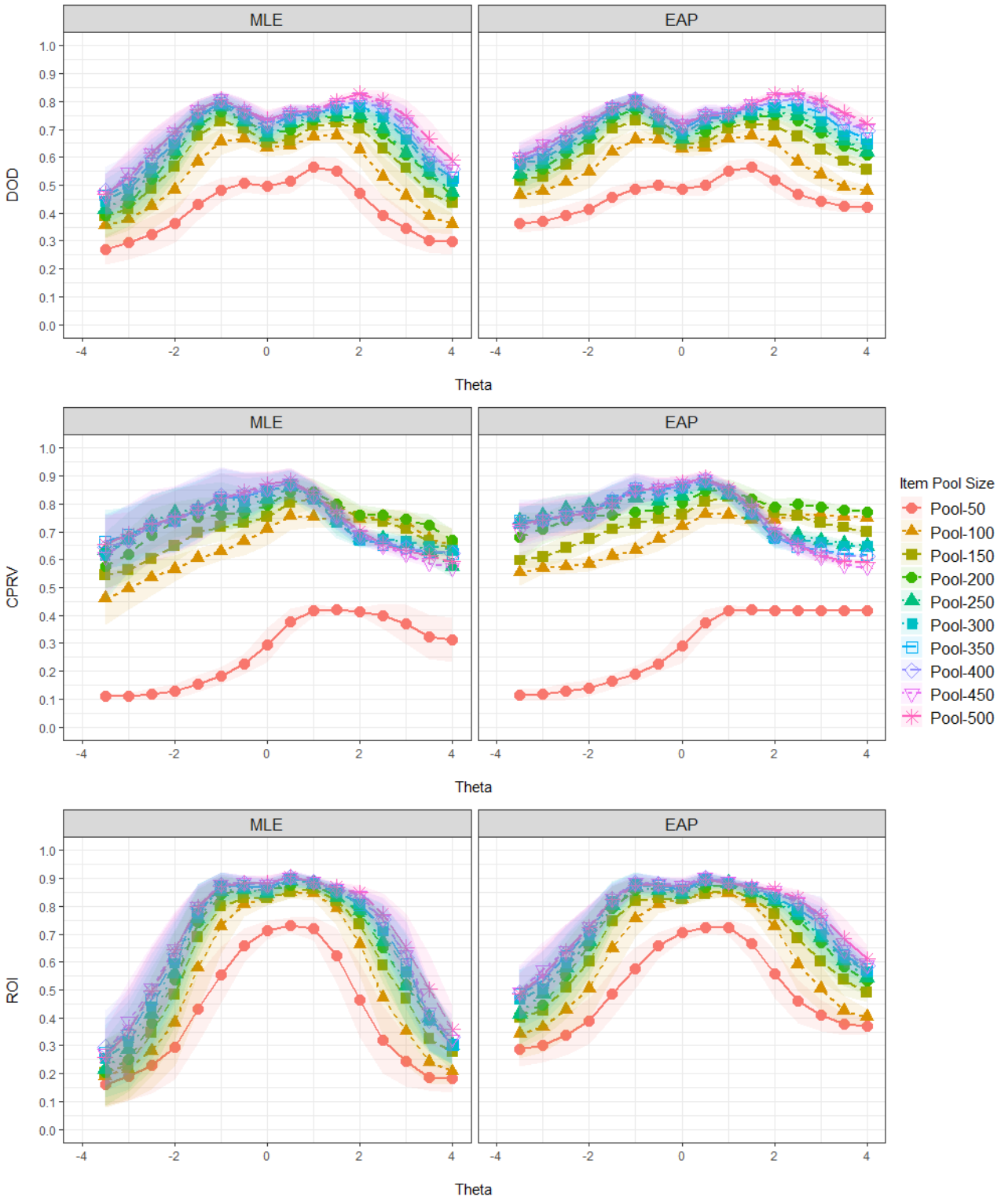


Figure 5.6. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a 3PL-based CAT by item pool size and proficiency estimator.

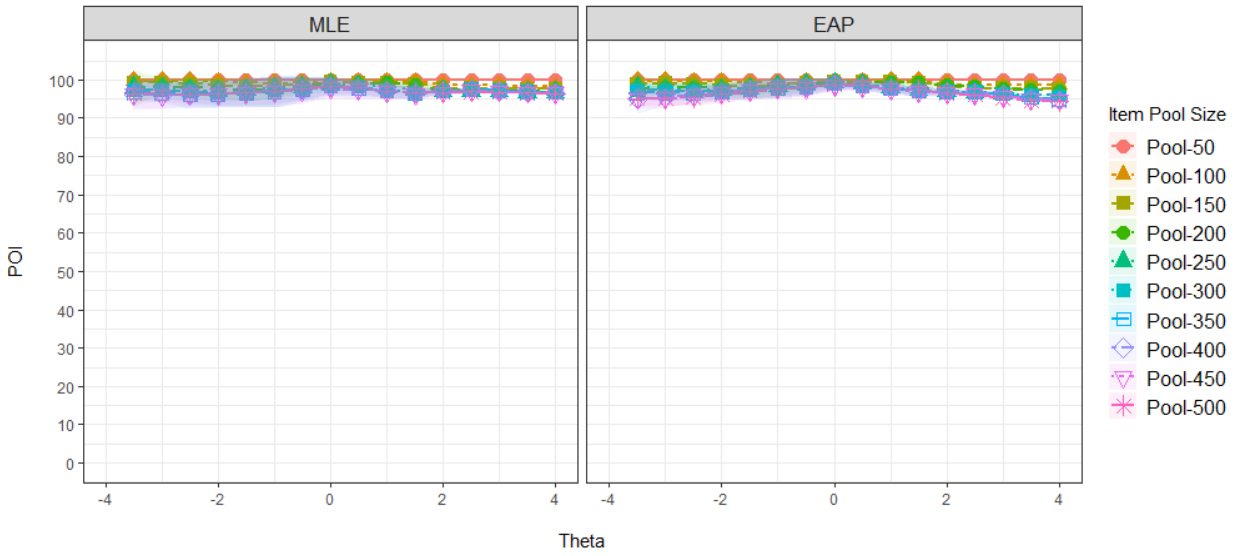


Figure 5.7. Plot of a POI index for a 3PL-based CAT by item pool size and proficiency estimator.

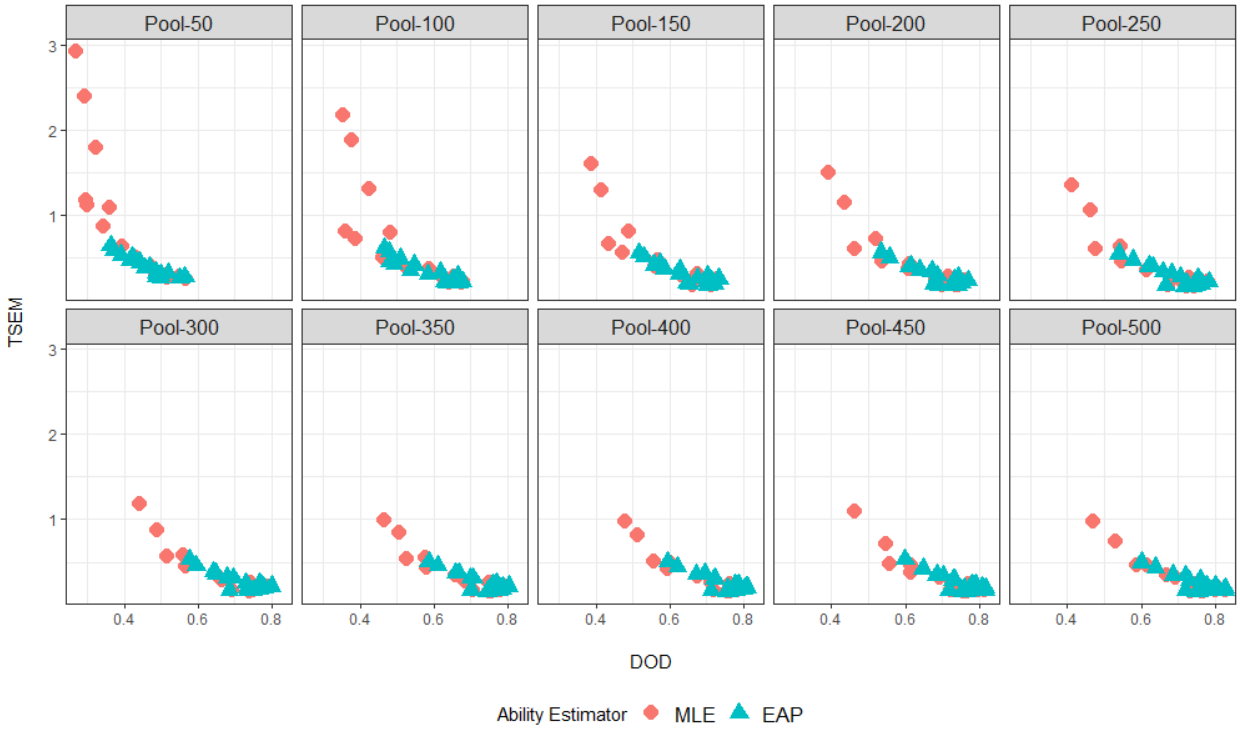


Figure 5.8. Relationship of TSEM with conditional adaptivity indices (DOD, CPRV, and ROI) for a 3PL-based CAT by item pool size and proficiency estimator.

Figure 5.8. (cont'd)

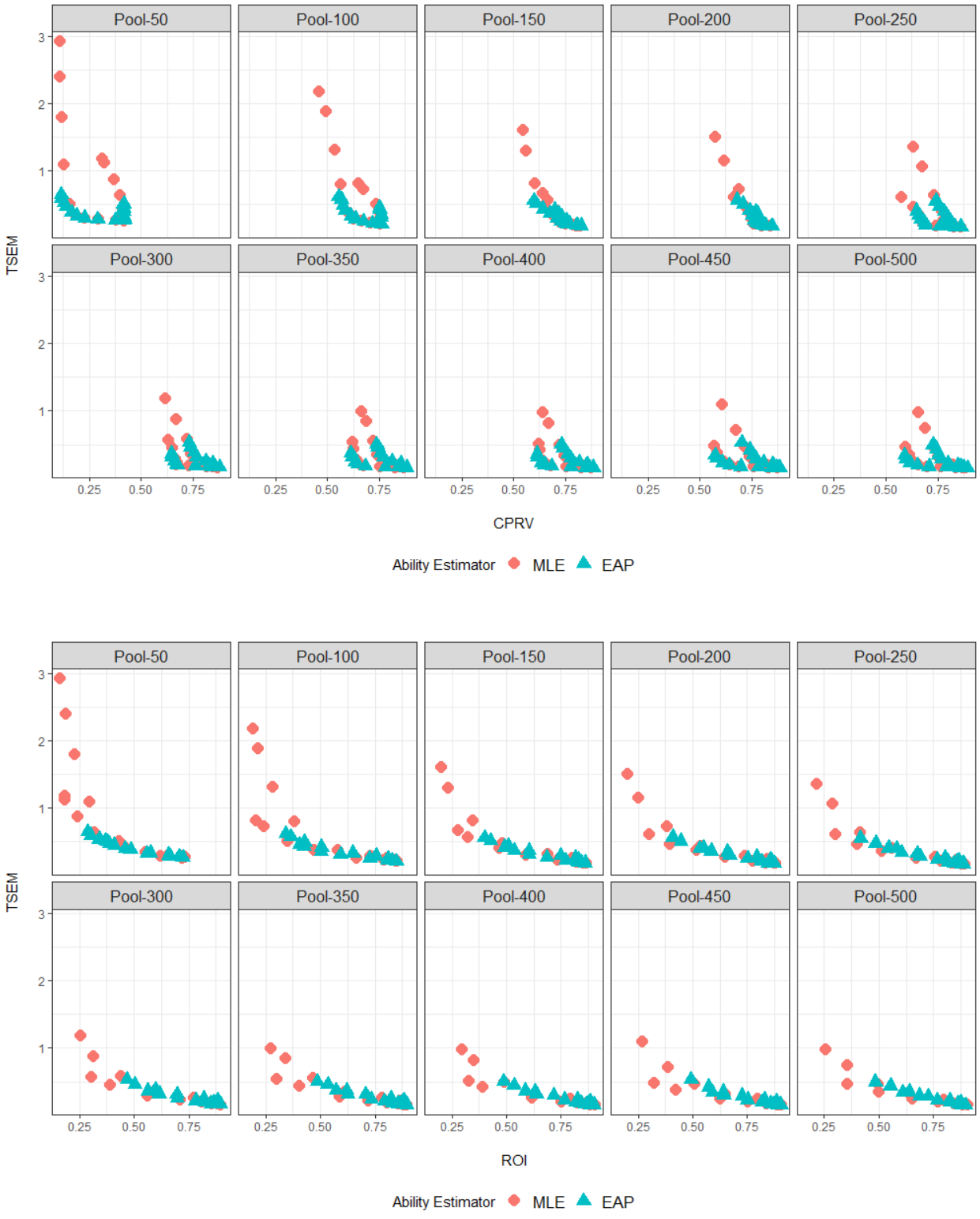


Table 5.4

Overall Adaptation Statistics for a 3PL-based CAT by Item Pool Size and Proficiency Estimator

Pool Size	MLE				EAP			
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI
	M (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
50	0.83 (.004)	0.25 (.002)	0.30 (.001)	99.92 (.007)	0.86 (.003)	0.28 (.002)	0.30 (.001)	99.95 (.005)
100	0.93 (.004)	0.55 (.005)	0.70 (.001)	99.34 (.027)	0.96 (.001)	0.59 (.003)	0.71 (.001)	99.48 (.021)
150	0.95 (.003)	0.66 (.006)	0.76 (.001)	98.98 (.033)	0.97 (.001)	0.69 (.003)	0.77 (.001)	99.16 (.021)
200	0.96 (.003)	0.70 (.004)	0.79 (.001)	98.54 (.041)	0.98 (.001)	0.73 (.003)	0.81 (.001)	98.82 (.034)
250	0.97 (.003)	0.72 (.005)	0.80 (.001)	98.12 (.051)	0.98 (.001)	0.74 (.003)	0.82 (.001)	98.51 (.033)
300	0.97 (.002)	0.74 (.004)	0.81 (.002)	97.68 (.060)	0.98 (.001)	0.75 (.002)	0.83 (.001)	98.23 (.042)
350	0.97 (.002)	0.75 (.004)	0.82 (.002)	97.44 (.062)	0.99 (.001)	0.76 (.003)	0.84 (.001)	98.07 (.041)
400	0.98 (.002)	0.77 (.003)	0.82 (.002)	97.29 (.059)	0.99 (.001)	0.78 (.003)	0.84 (.001)	97.90 (.038)
450	0.98 (.002)	0.78 (.004)	0.82 (.002)	97.16 (.070)	0.99 (.001)	0.79 (.003)	0.84 (.001)	97.81 (.045)
500	0.98 (.001)	0.78 (.004)	0.83 (.002)	97.49 (.062)	0.99 (.001)	0.79 (.002)	0.85 (.001)	97.94 (.040)

Note. All statistics were computed using final estimated proficiencies (θ).

5.1.2 Variation in item pool spread

Another characteristic of item pools that could affect the amount of adaptation is the magnitude of the spread of item characteristics (*b*-parameter or location of maximum information) in the item pool. The second section of Research Question 1 aimed to examine the sensitivity of three conditional adaptation indices to variations in item pool spread. It was hypothesized that if the difficulty of the items in the pool is in a limited range, even though the item pool is large, the CAT cannot be suitably customized for students whose proficiency levels are outside that range covered by the item pool. To test the hypothesis, eight item pools were simulated that differed in the *SD*s of *b*-parameters in an item pool from 0.2 to 1.6 at 0.2 intervals. The size of all item pools considered here was 400. In the following, the impact of item pool spread was investigated moderated by IRT models (Rasch and 3PL) and proficiency estimators (MLE and EAP) in terms of measurement accuracy and precision as well as the amount of adaptation.

5.1.2.1 Rasch model

Measurement accuracy and precision. The final proficiency estimates were evaluated using three conditional- and four overall statistics for measurement accuracy and precision. A smaller bias value indicates a more accurate proficiency estimate, and a smaller TSEM value indicates a more precise and stable proficiency estimate. The RMSE considers both bias and standard errors together so that a smaller RMSE is associated with better recovery of proficiency estimates. Also, the higher correlation between true and final proficiency estimates ($r_{\theta\hat{\theta}}$) is associated with better recovery.

Conditional statistics. As shown in Figure 5.9, MLE presented little bias scattering the mean proficiency estimates around 0.0, while EAP showed biased estimates regressed toward the

mean of a prior distribution (i.e., 0.0) on the proficiency scale. These findings were consistent across all eight item pools that had different SDs of *b*-parameters, though the limited item pool with small *SD*s of *b*-parameters showed slightly more bias in the estimates. However, MLE provided larger standard errors (TSEM) for the proficiency estimates than EAP, especially at the extreme regions of the proficiency continuum. As the item pool spread was restricted, TSEM increased, implying less precision in the proficiency estimates. Overall, the RMSE values of the two estimators were similar to each other at the moderate proficiency levels, whereas they were obviously greater for EAP at the extreme positive and negative ends of the proficiency scale.

Overall statistics. Table 5.5 presents the summary of four overall statistics by variation in item pool spread and proficiency estimator. In general, slightly more bias was found in the estimates using EAP, while larger standard errors (TSEM) were identified using MLE over the entire sets of data. The differences either in bias or in standard errors between the two estimators were small, becoming negligible with larger item pools. The correlation coefficients, $r_{\theta\hat{\theta}}$, were almost identical between the two estimators, and the degree of the correlation improved as the *SD*s of *b*-parameters increased in the item pool. More interestingly, regardless of variations in item pool spread, the RMSE was smaller for EAP than for MLE, implying that the final proficiency estimates were more accurately, precisely measured using EAP, on average, over the entire sample of students.

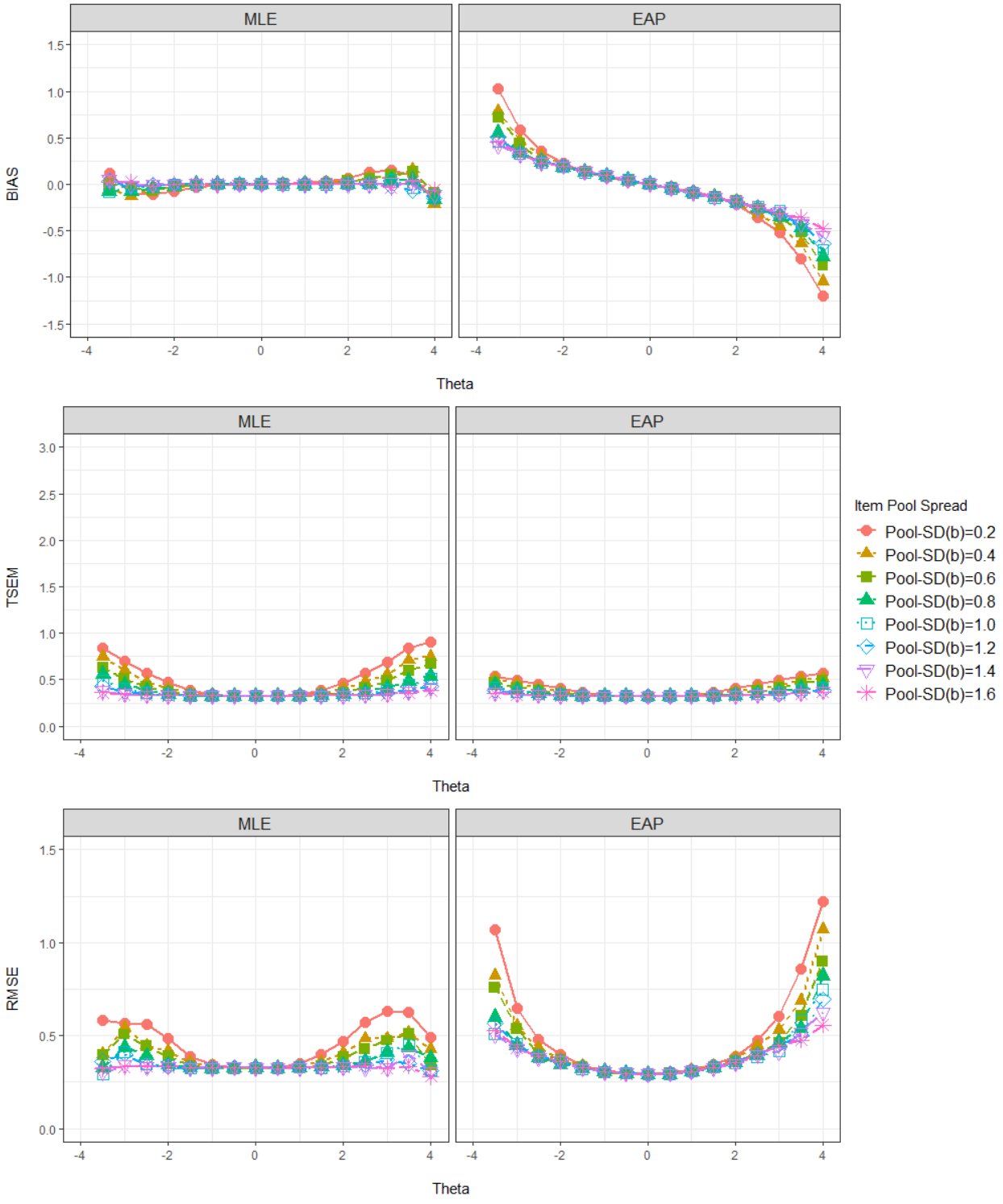


Figure 5.9. Conditional bias, TSEM, and RMSE of proficiency estimates for a Rasch-based CAT by item pool spread and proficiency estimator.

Table 5.5

Overall Statistics of Measurement Precision of Proficiency Estimates for a Rasch-based CAT by Item Pool Spread and Proficiency Estimator

Pool	MLE				EAP			
SD (*b*s)	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
0.2	0.002	0.356	0.361	0.945	-0.002	0.341	0.325	0.947
0.4	0.001	0.340	0.343	0.948	-0.003	0.331	0.318	0.949
0.6	0.001	0.332	0.336	0.950	-0.002	0.326	0.314	0.950
0.8	0.000	0.329	0.331	0.950	-0.002	0.324	0.311	0.951
1.0	-0.001	0.327	0.330	0.951	-0.003	0.323	0.311	0.951
1.2	0.000	0.326	0.330	0.950	-0.002	0.323	0.311	0.951
1.4	0.001	0.326	0.330	0.950	-0.002	0.323	0.311	0.951
1.6	0.000	0.326	0.329	0.951	-0.002	0.323	0.310	0.951

Note. *b*s = *b*-parameters.

Amount of adaptation. The proposed conditional adaptivity indices, along with the overall indices, were used to assess the difference in adaptivity for the CATs.

Conditional adaptivity. Figure 5.10 presents the distributions of three conditional adaptation indices over the proficiency continuum using eight item pools that varied in the *SD*s of *b*-parameters by two proficiency estimators. Overall, the three adaptivity measures sensitively detected each corresponding aspect of the amount of adaptation for the CATs depending on the extent of the item pool spread. The proposed statistics generally increased as the *b*-parameters were more broadly spread out in the pool. In particular, at the extreme regions of the proficiency continuum, it was clearly observed that the values of the three indices gradually improved as the item pool contained more difficult or easy items. For the item pool with 1.6 *SD* of the *b*-parameters, the three measures indicated that the CATs were almost equally well adapted across all proficiency levels. Looking at the performance of each index, the DOD and ROI indices functioned as expected depending on variation in item pool spread over the proficiency

continuum; however, for the CPRV index, an unexpected pattern was observed when the *SD* of *b*-parameters in the item pool was very small (i.e., 0.2 or 0.4). The CPRV values using those item pools were exceptionally small, closer to 0.0 or even below 0.0 on the moderate proficiency levels. Since the CPRV index compares the variation of the *b*-parameters of the items selected for each examinee relative to the variation of *b*-parameters in the entire item pool, if the item pool includes most items whose difficulty were in a very restricted range, say 0.04 or 0.16 variances, even if the within-examinee variance is per se small, that variance could be larger *relative* to the item pool variance. Note that as with the findings for the item-pool-size study, the POI index showed little sensitivity across the proficiency continuum and variation in the spread of *b*-parameters in the item pool (see Figure 5.11).

The three adaptation measures showed similar patterns between the MLE and the EAP using the eight item pools, but their observed ranges of the values over the proficiency continuum were different with broader ranges for the MLE. That is, compared to the MLE, all three adaptation measures presented less variations in the corresponding values across the proficiency levels for the CATs using EAP, which was consistent across the eight item pools. Again, this might due to the features of the EAP estimator yielding the regress-toward-the-mean bias in the estimates.

Regarding the stability of the indices, the DOD and ROI measures reported small empirical standard errors over the proficiency scales regardless of variations in item pool spread. However, the CPRV index showed poor stability when the *b*-parameters were in a very restricted range in the pool, although the index was stable with the item pools that had the *SD* of *b*-parameters larger than 0.8. This instability was even greater when the MLE was used for the CATs because a fixed-step size of 0.7 was used earlier in the test until the MLE can be

computed. This suggests that given that students' abilities followed the standard normal distribution, N (0, 1), the variation of b-parameters for an item pool should be about equal to or greater than the variation in the final proficiency estimates in order for the CPRV index to be precise.

Finally, these findings for the item-pool-spread study support the benchmark values of the proposed conditional indices, suggested in the previous item-pool-size study in Section 5.1.1, when the Rasch model is used for the CATs. For the 40-item test, mid 0.90s for DOD, high 0.70s for CPRV, and high 0.90s for ROI indicate good adaption using the MLE, while using the EAP, high 0.90s for DOD, high 0.80s for CPRV, and high 0.90s for ROI considered good adaptation.

Overall adaptivity. The results of overall adaptivity for the item pool spread showed that the correlation, ratio of SDs, and PRV indices gradually improved as the spread of the item pool difficulty increased (see Table 5.6). Compared to the MLE, using the EAP reported larger values of the correlation and PRV measures but smaller values for the ratio of SDs index. As mentioned earlier, the latter is attributed to the property of the EAP estimator. However, as with the results for the item pool size study, the POI index was rarely sensitive to the spread of the item pool difficulty, and its value decreased as the spread of the item pool increased. Overall, these results were in line with the item-pool-size study when selecting benchmark values for the measures. At the same time, it can be confirmed that based on the benchmark values, the variation of b-parameters for an item pool should be larger than the variation in proficiency estimates for the CATs to be well adapted to students whose proficiency is at the extremes of the proficiency scale.

Relationship between conditional adaptivity and measurement precision. Figure 5.12 displays the relations of the standard errors (TSEM) with the DOD, CPRV, and ROI

indices. The DOD and ROI measures were negatively associated with the standard errors with a linear relationship but with a slightly nonlinear relation between DOD and TSEM for MLE with the limited spread of the item pool difficulty. The Pearson correlation coefficients were also computed, showing high, strong correlation with one another. The DOD and TSEM correlations were in the range of -0.98 to -0.90 for EAP as well as in -0.98 to -0.74 for MLE. The ROI and TSEM correlations were in the range of -0.98 to -0.81 for EAP as well as -0.97 to -0.81 for MLE. However, for the relation with CPRV, it is interesting to note that on the one hand, when the *SD* of *b*-parameters in the pool was smaller than 0.6, the CPRV values were in general positively correlated with the TSEM values; on the other hand, when the spread of *b*-parameters was equal or greater than 0.8, their relationship appeared to be negative. Again, the former can be explained by the unusual pattern identified in Figure 5.10 with the restricted spread of the item pool. Also, when the spread of difficulty in the pool was large, the relation between CPRV and TSEM looked linear.

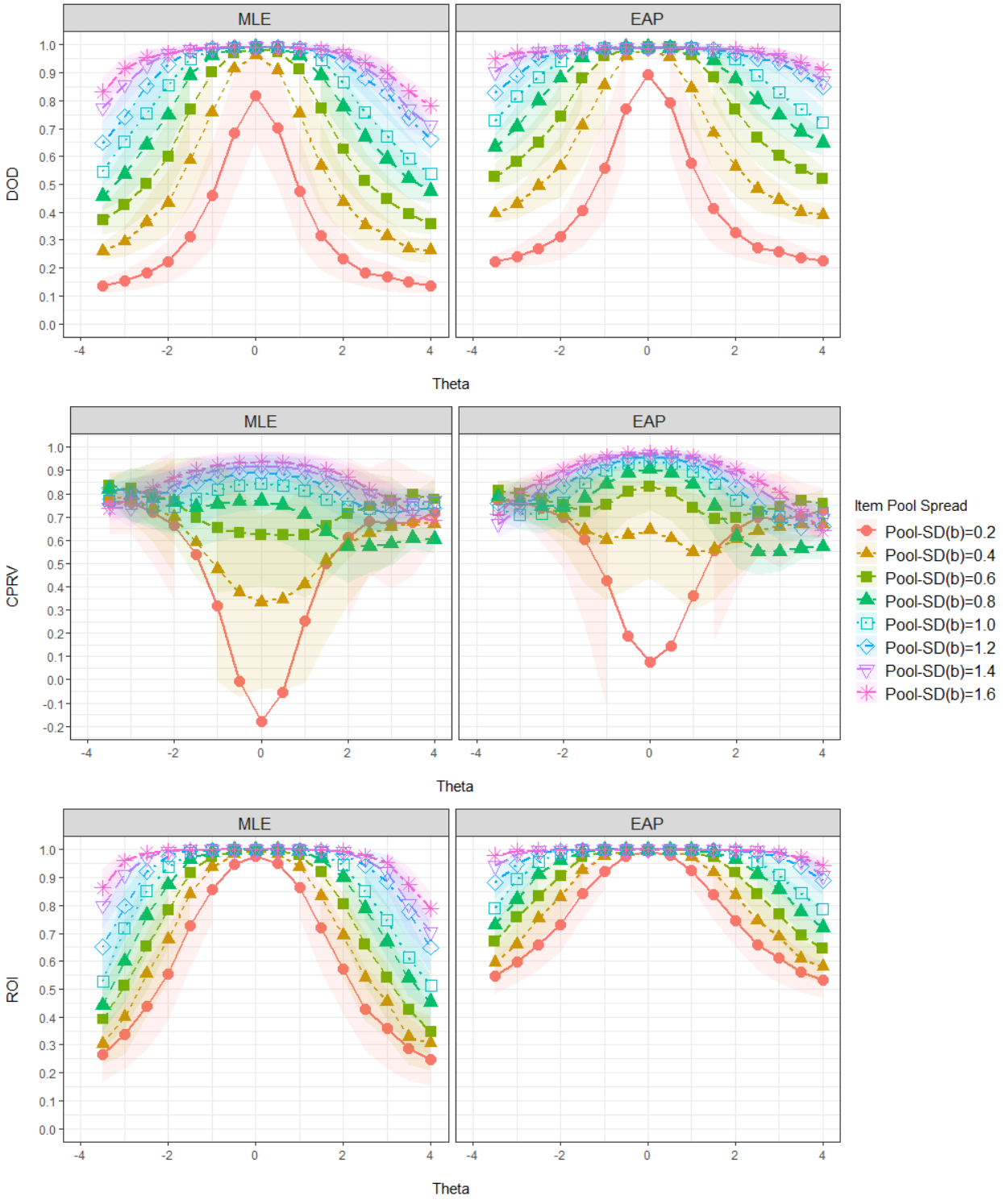


Figure 5.10. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a Rasch-based CAT by item pool spread and proficiency estimator.

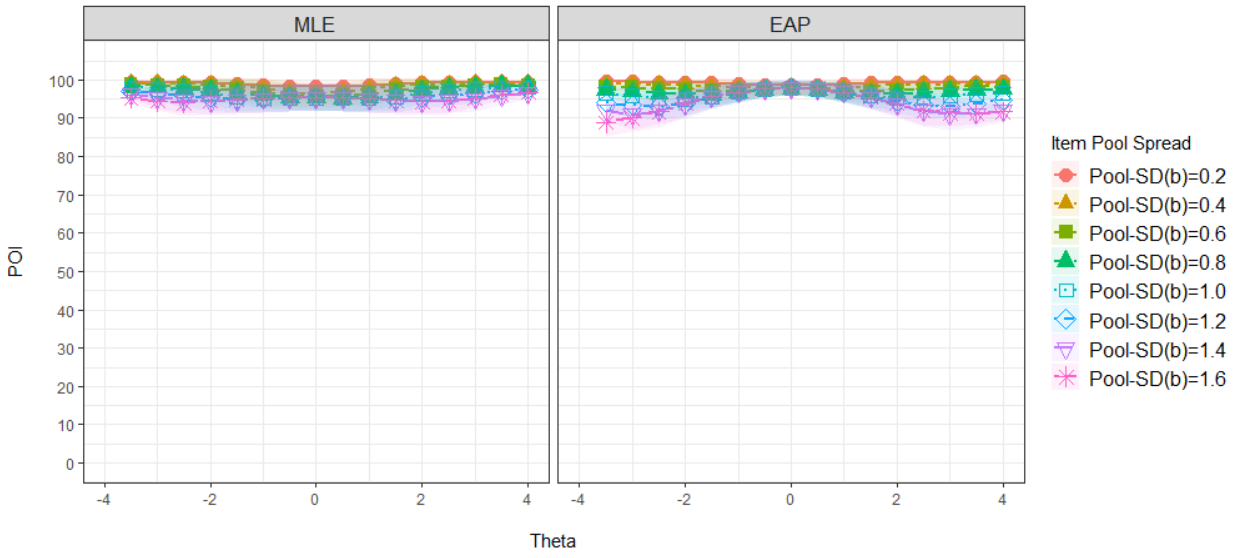


Figure 5.11. Plot of a POI index for a Rasch-based CAT by item pool spread and proficiency estimator.

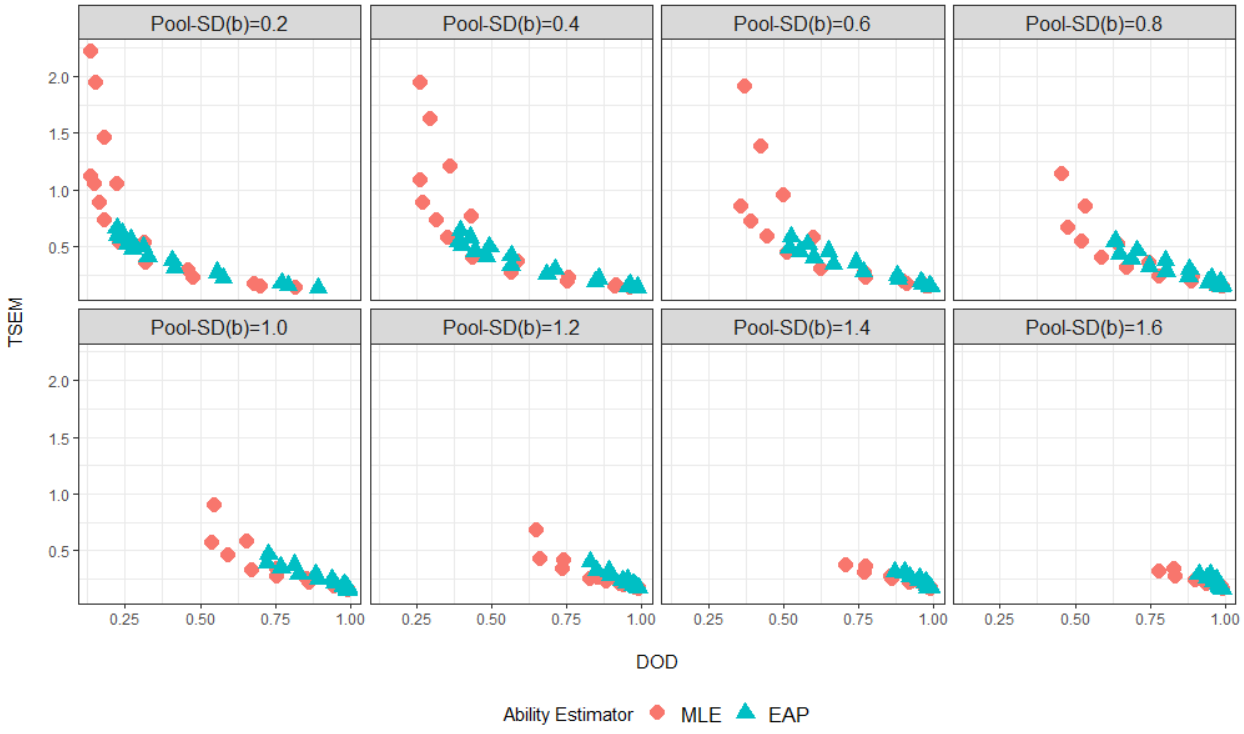


Figure 5.12. Relationship of TSEM with conditional adaptivity indices for a Rasch-based CAT by item pool spread and proficiency estimator.

Figure 5.12. (cont'd)

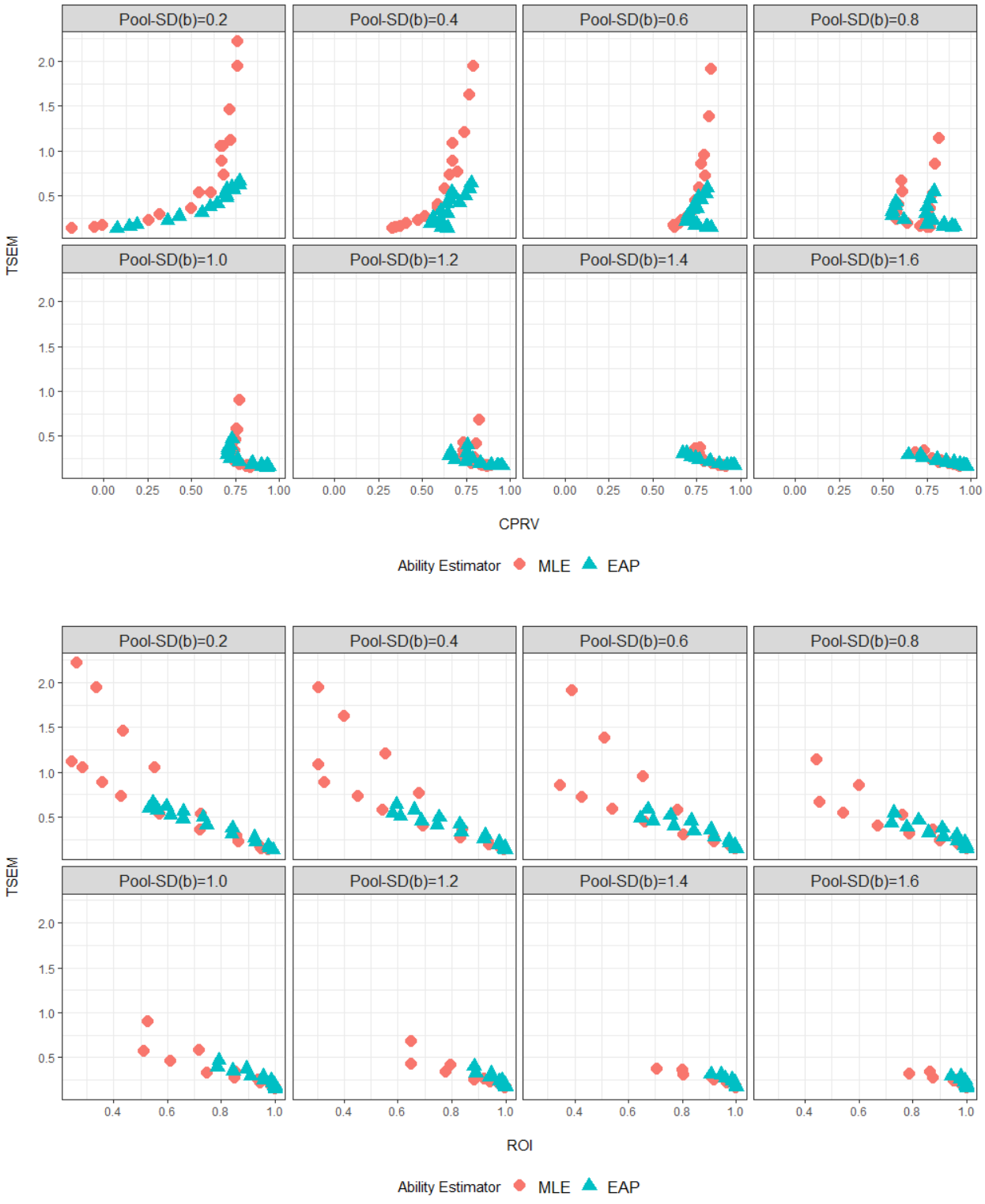


Table 5.6

Overall Adaptation Statistics for a Rasch-based CAT by Item Pool Spread and Proficiency Estimator

Pool *SD* (*b*s)	MLE				EAP			
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI
	M (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
0.2	0.84 (.003)	0.26 (.003)	0.15 (.016)	98.61 (.034)	0.87 (.003)	0.30 (.002)	0.30 (.012)	98.82 (.027)
0.4	0.88 (.003)	0.50 (.004)	0.43 (.011)	97.30 (.064)	0.91 (.002)	0.53 (.004)	0.61 (.006)	97.97 (.053)
0.6	0.91 (.003)	0.68 (.004)	0.65 (.005)	96.45 (.066)	0.94 (.002)	0.68 (.004)	0.78 (.003)	97.41 (.052)
0.8	0.93 (.002)	0.81 (.004)	0.74 (.004)	95.78 (.087)	0.96 (.001)	0.77 (.004)	0.85 (.002)	97.13 (.056)
1.0	0.94 (.002)	0.88 (.006)	0.82 (.002)	95.42 (.076)	0.97 (.001)	0.80 (.004)	0.90 (.001)	96.94 (.048)
1.2	0.95 (.002)	0.93 (.006)	0.87 (.002)	95.18 (.082)	0.97 (.001)	0.83 (.005)	0.93 (.001)	96.82 (.056)
1.4	0.96 (.001)	0.96 (.006)	0.90 (.001)	95.10 (.078)	0.97 (.001)	0.83 (.004)	0.95 (.001)	96.82 (.052)
1.6	0.96 (.001)	0.97 (.005)	0.92 (.001)	95.13 (.073)	0.97 (.001)	0.83 (.004)	0.96 (.001)	96.88 (.049)

Note. All statistics were computed using final estimated proficiencies (θ).

5.1.2.2 3PL model

Measurement accuracy and precision. The final proficiency estimates were evaluated using three conditional- and four overall statistics for measurement accuracy and precision.

Conditional statistics. As shown in Figure 5.9, using the 3PL model, the MLE yielded small bias when the restricted spread of the item pool was used, while the EAP reported bias in the estimates regressed toward the mean of a prior distribution over the proficiency continuum regardless of the spread of the item pools. The extent of bias became smaller with the bigger spread of the item pools. The estimates for the MLE were underestimated while those for the EAP were overestimated at the negative extreme region of the proficiency scale, and vice versa at the positive extreme proficiency region. Meanwhile, as with the results for the Rasch model, the MLE produced larger standard errors (TSEM) in the proficiency estimates compared to the EAP especially at the extremes of the proficiency scale. As the *SD* of *b*-parameters in the item pool increased, the standard errors reduced, indicating more precision in the estimates. Considering both bias and standard errors, the RMSE values were small to a similar extent for the two estimators at the middle proficiency levels around 0.0, whereas they were greater at the extremes of the proficiency scale and even the EAP presented higher values at the very ends of the scale than MLE did.

Overall statistics. In general, as the spread of the item pool increased, the standard errors (TSEM and RMSE) decreased and the correlation coefficients, $r_{\theta\hat{\theta}}$, improved. Although the correlations were similar between MLE and EAP, as similar to the results for the Rasch model, the grand means of standard errors were greater for MLE than for EAP. The differences became small as the *SD* of *b*-parameters increased, though. Again, it suggests that the final proficiency

estimates were more precisely measured using EAP on average over the entire group of students compared to MLE.

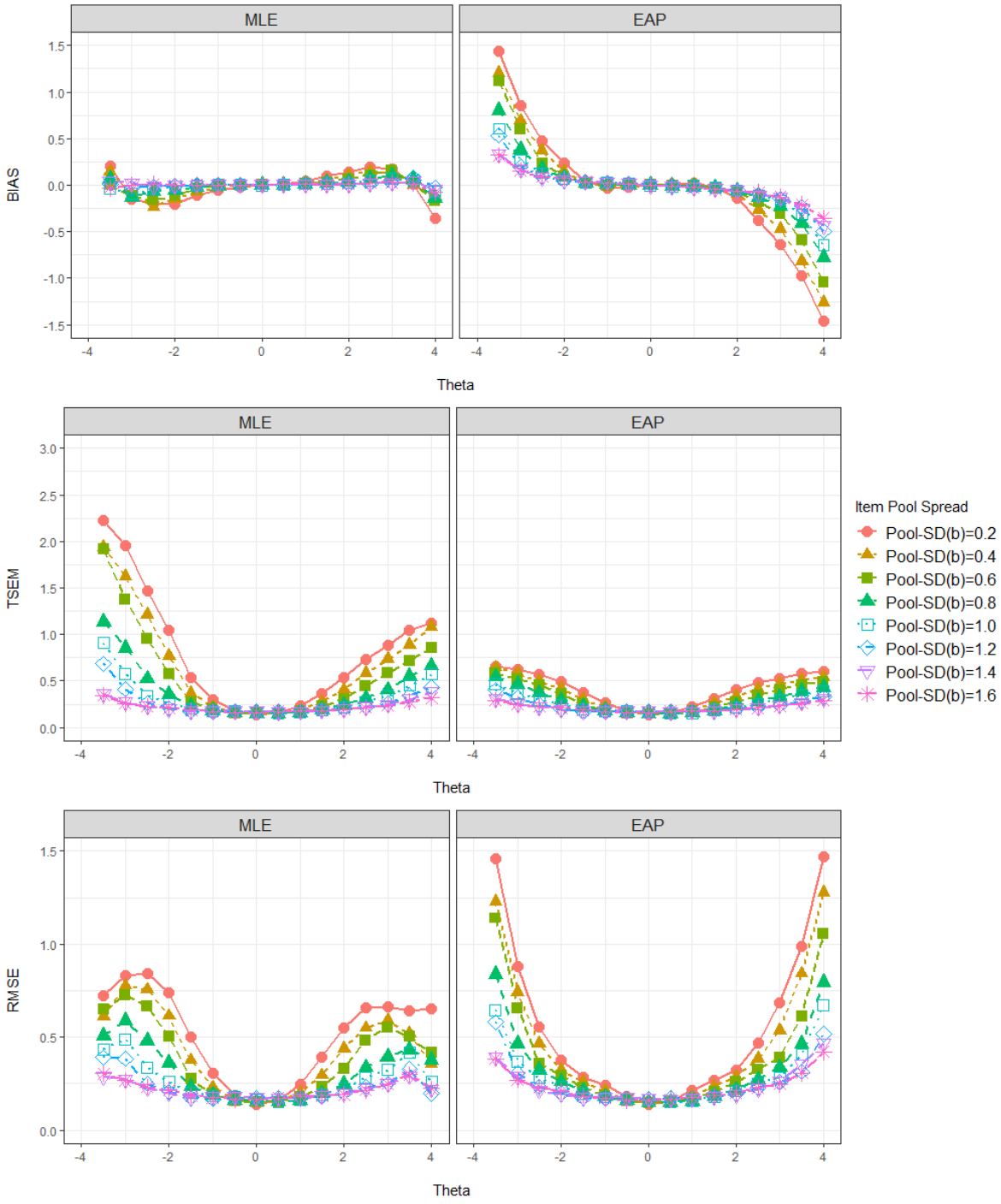


Figure 5.13. Conditional bias, TSEM, and RMSE of proficiency estimates for a 3PL-based CAT by item pool spread and proficiency estimator.

Table 5.7

Overall Statistics of Measurement Precision of Proficiency Estimates for a 3PL-based CAT by Item Pool Spread and Proficiency Estimator

Pool	MLE				EAP			
SD (*b*s)	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
0.2	-0.003	0.376	0.320	0.960	-0.001	0.253	0.234	0.973
0.4	-0.001	0.295	0.262	0.971	-0.001	0.218	0.208	0.978
0.6	0.000	0.244	0.225	0.978	-0.001	0.195	0.190	0.982
0.8	0.002	0.196	0.195	0.982	0.000	0.183	0.179	0.984
1.0	0.003	0.182	0.184	0.984	-0.001	0.177	0.175	0.985
1.2	0.004	0.178	0.180	0.985	-0.001	0.176	0.176	0.985
1.4	0.005	0.177	0.180	0.985	-0.001	0.176	0.176	0.985
1.6	0.005	0.175	0.178	0.985	-0.001	0.174	0.174	0.985

Note. *b*s = *b*-parameters.

Amount of adaptation. Adaptivity of the CATs was evaluated using the conditional measures at the individual proficiency level and the overall indices over the entire sample of students.

Conditional adaptivity. Results for the effects of variation in item pool spread using the 3PL model indicated that the three adaptation measures (DOD, CPRV, and ROI) appeared to be sensitive to the spread of the item pool across the examinees' proficiency levels (see Figure 5.14). The proposed statistics mostly increased as the difficulty parameters were more broadly spread out in the pool. In particular, at the extreme regions of the proficiency continuum, it was observed that the values of the three indices gradually improved as the item pool contained more difficult or easy items. Given the assumed standard normal distribution of examinee's proficiency and the 40-item test, these results suggested that the variation of *b*-parameters in an item pool should be larger than the variation of examinees' proficiency in order to achieve good adaptivity over the examinees, especially for those at the extremes of proficiency. However, as

found in the previous investigations, the POI index appeared to be insensitive to the spread of items in the pool using the 3PL model, though the POI values were slightly lower at the extremes of the proficiency scale regardless of the proficiency estimators (see Figure 5.15). Considering the concept of the index, it might be expected that given the available item pool, the optimal test information would be similar to the observed test information unless the interim estimate deviated far from the final estimate or the CAT included many constraints on the item selection.

Compared to the results for the Rasch model in Section 5.1.2.1, even though the conditional adaptation statistics were computed using the location of maximum information (Birnbaum, 1968) instead of *b*-parameters, the values of the three indices were relatively lower over the proficiency continuum when the 3PL model was used for the CATs. Also, the unusual pattern was not identified in the plot of CPRV for the 0.2 *SD* of *b*-parameters for the item pool condition. These might be due to the effect of *a*- and *c*-parameters on proficiency estimation as well as on the information function, affecting the item selection procedure for the CATs using 3PL model. However, the pattern of DOD was slightly different from the DOD's pattern using the Rasch model because of the characteristics of the restricted item pool interacted with the effect of *a*- and *c*-parameters. With respect to the comparison between MLE and EAP, the three statistics showed similar patterns between the two proficiency estimators across the eight item pools, but their observed values were generally greater with the smaller ranges over the proficiency continuum for the EAP.

Regarding the stability of the three measures, as with the results of the variation-in-pool-size study using the 3PL model, the three statistics had greater empirical standard errors at the very top and bottom ends of the proficiency continuum rather than those at the moderate

proficiency levels. As mentioned before, this may be due to the effect of *c*-parameters on the proficiency estimates for the 3PL model.

In sum, these results were consistent with the findings for the item pool size study when the 3PL model was used for the CATs, supporting the benchmark values for the three statistics to indicate good adaptation: For MLE, a value in the mid 0.70's for DOD, a value in the low 0.80's for CPRV, and a value in the mid 0.80's for ROI. For EAP, a value in the high 0.70's for DOD, a value in the mid 0.80's for CPRV, and a value in the mid 0.80's for ROI.

Overall adaptivity. Similar to prior studies using the 3PL model (Reckase et al., 2018; Ju & Lee, 2018), as the spread of the *b*-parameters for the item pool increased, the overall adaptivity indices gradually improved with the most sensitive of the ratio of standard deviations index (see Table 5.8). These results gave additional evidence to support the statistics selected for indicating good adaptation over the entire group of students using the overall measures: For MLE, a value in the high 0.90's for the correlation, a value in the high 0.70's for the ratio of SDs, and a value in the low 0.80's for PRV. For EAP, a value in the high 0.90's for the correlation, a value in the high 0.70's for the ratio of SDs, and a value in the mid 0.80's for PRV.

Relationship between conditional adaptivity and measurement precision. As seen in Figure 5.16, except for the very large spread of the item pool whose *SD* of *b*-parameters was greater than 1.2, the DOD measure was negatively, strongly correlated with TSEM ($-.93 < r\text{s} < -0.80$ for MLE; $-0.98 < r\text{s} < -0.73$ for EAP). While the relationship between ROI and TSEM was slightly nonlinear for the very restricted spread of the item pool, they were negatively and very closely associated with one another for other pool spread conditions ($-0.99 < r\text{s} < -.83$ for MLE; $-0.99 < r\text{s} < -.81$ for EAP). This was expected because both TSEM and ROI were computed using the information. Lastly, for the relation of CPRV, it was hard to find a systematic pattern of the

relation across the item pool spread conditions, but with the high spread of the item pool whose *b*-parameters were greater than 1.0, regardless of the proficiency estimators, TSEM had a negative, strong, and linear relation with CPRV ($-0.93 < r\text{s} < -.74$ for MLE; $-0.97 < r\text{s} < -0.68$ for EAP).

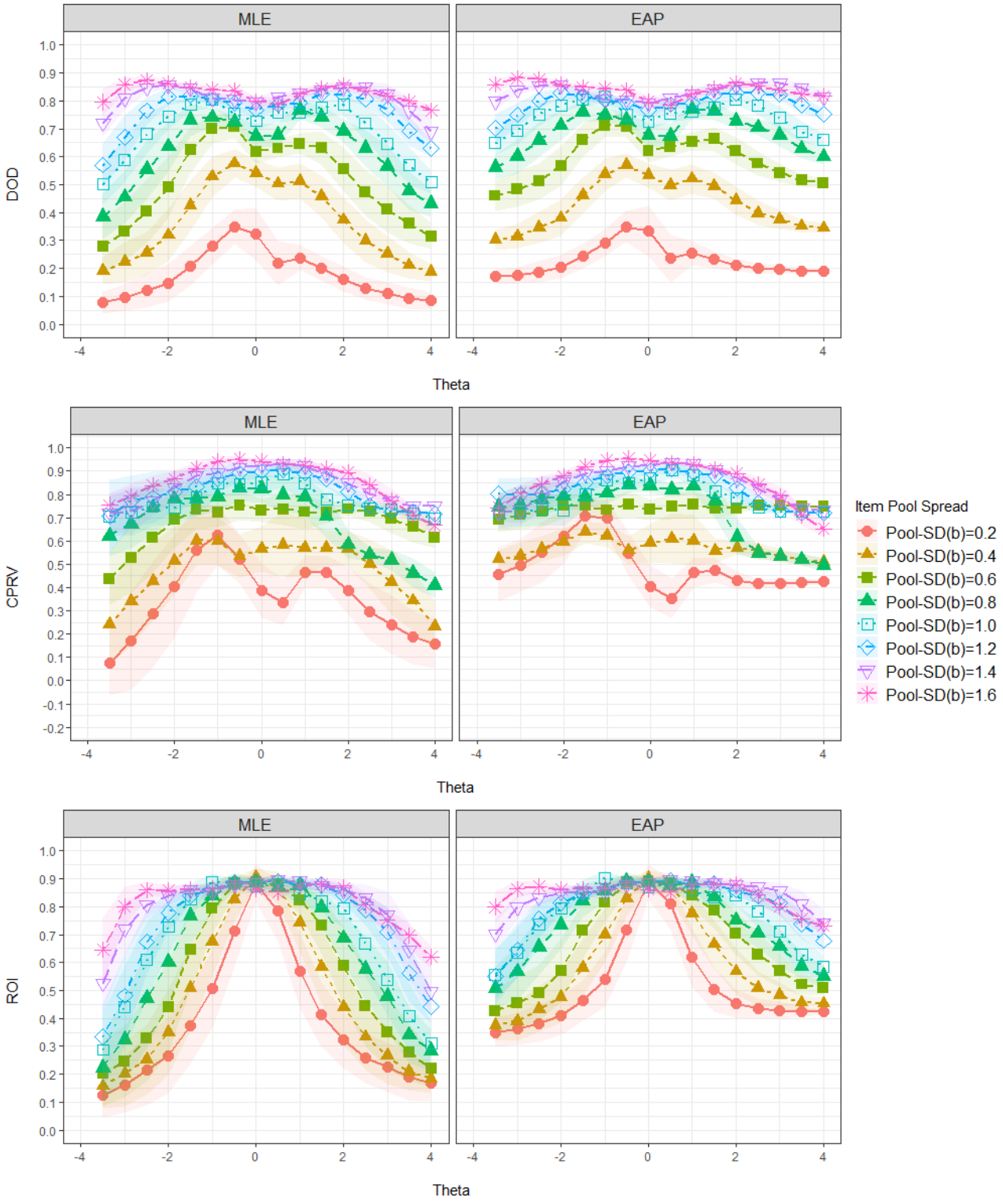


Figure 5.14. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a 3PL-based CAT by item pool spread and proficiency estimator.

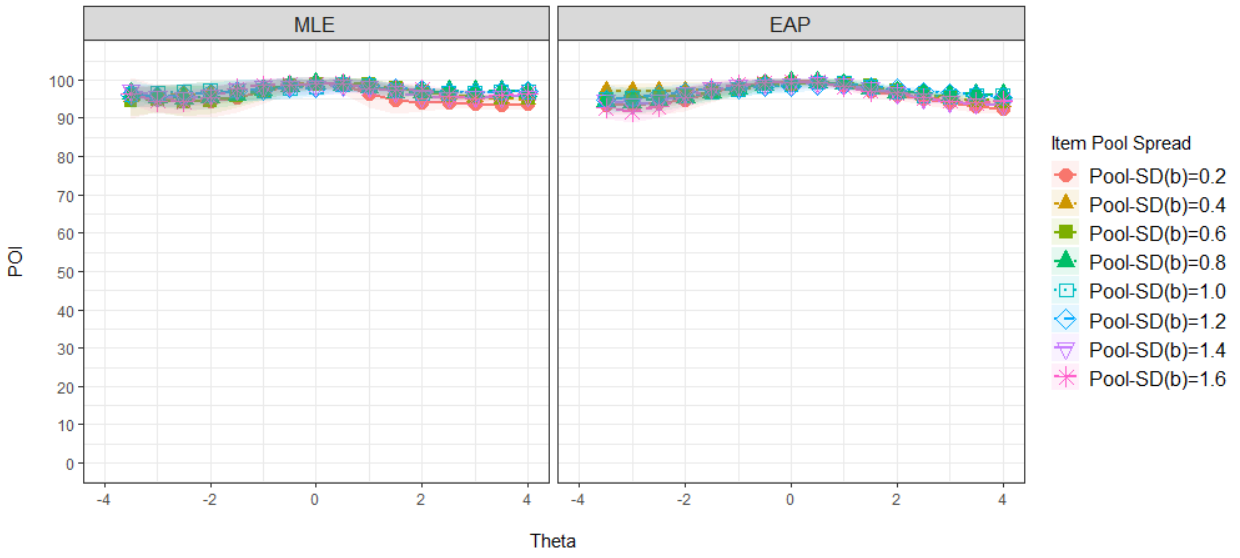


Figure 5.15. Plot of a POI index for a 3PL-based CAT by item pool spread and proficiency estimator.

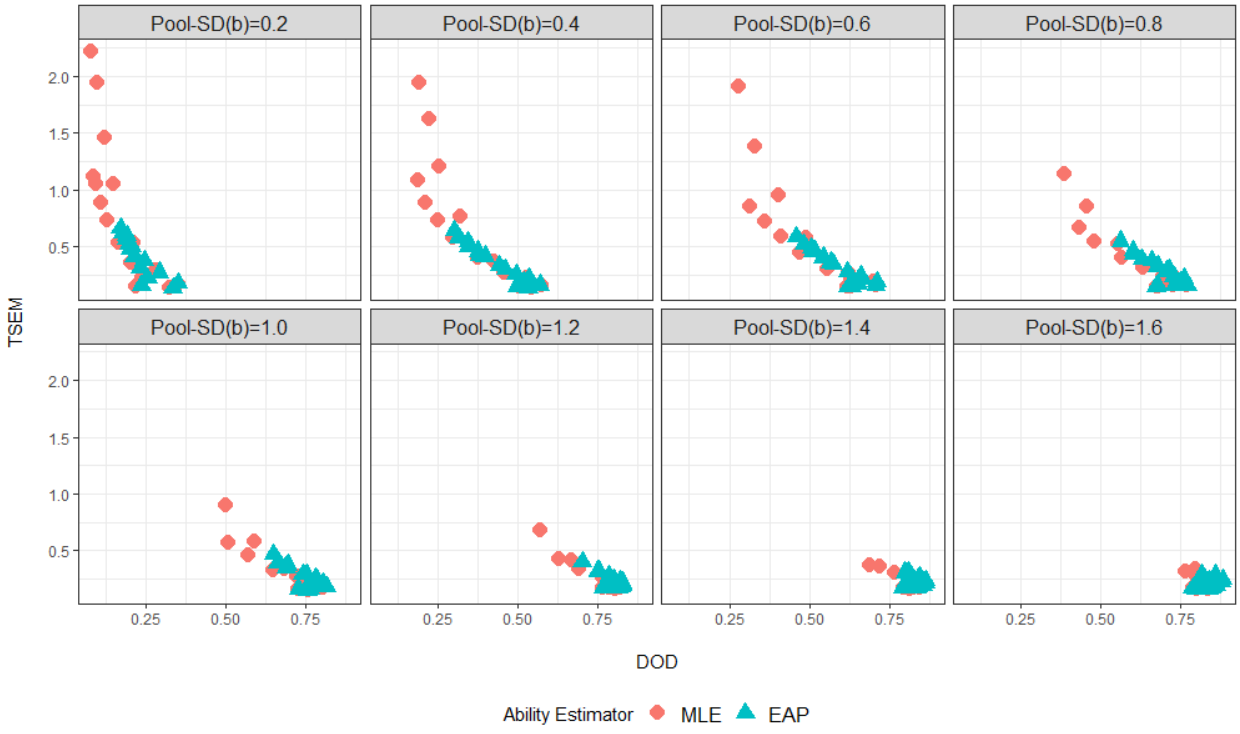


Figure 5.16. Relationship of TSEM with conditional adaptivity indices (DOD, CRPV, and ROI) for a 3PL-based CAT by item pool spread and proficiency estimator.

Figure 5.16. (cont'd)

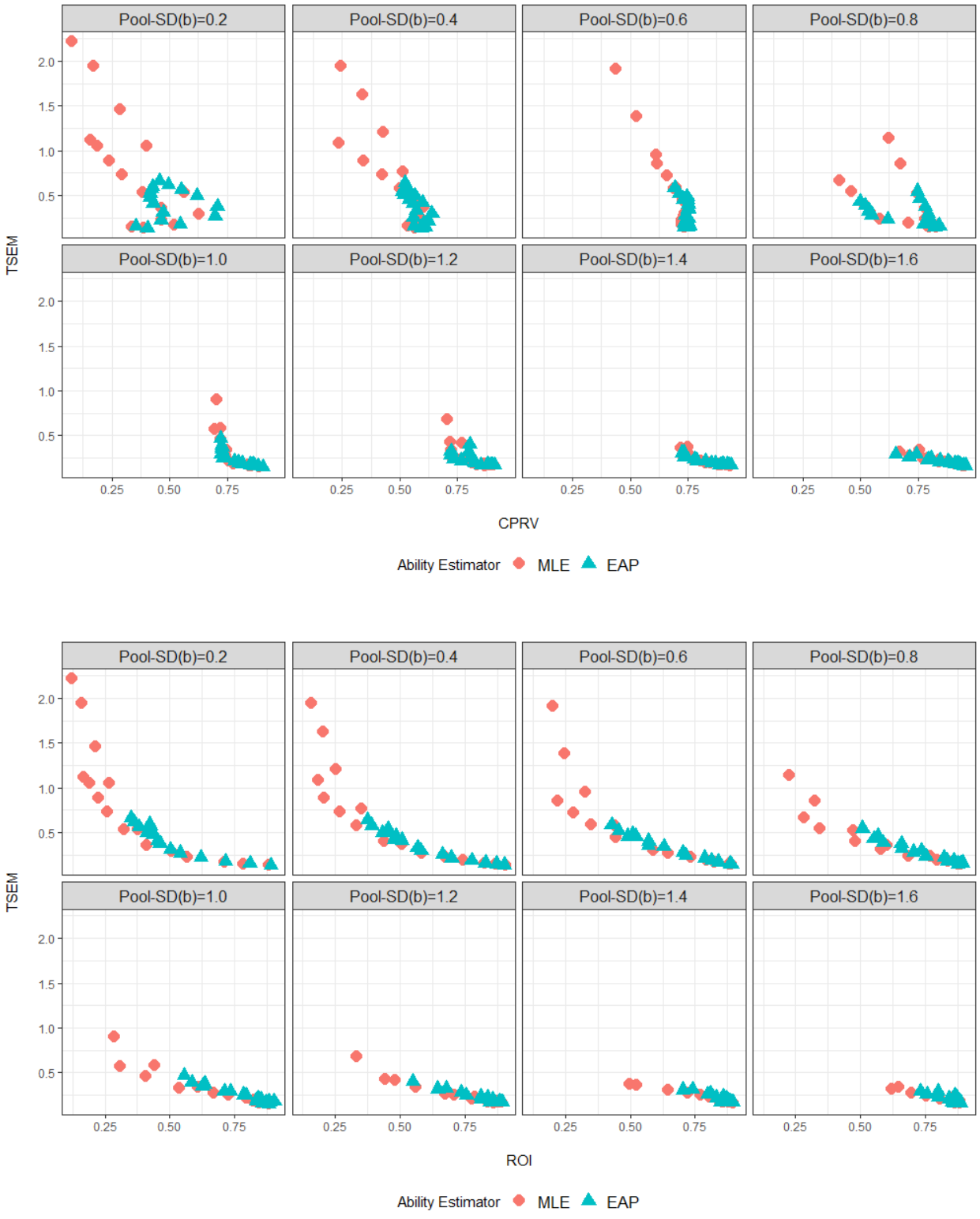


Table 5.8

Overall Adaptation Statistics for a 3PL-based CAT by Item Pool Spread and Proficiency Estimator

Pool *SD* (*b*s)	MLE				EAP			
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV	POI
	M (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
0.2	0.87 (.006)	0.18 (.002)	0.45 (.003)	97.43 (.065)	0.96 (.002)	0.21 (.001)	0.49 (.002)	98.55 (.025)
0.4	0.92 (.006)	0.38 (.004)	0.57 (.002)	98.24 (.040)	0.96 (.001)	0.42 (.002)	0.59 (.002)	98.59 (.027)
0.6	0.95 (.005)	0.56 (.005)	0.73 (.001)	98.04 (.038)	0.98 (.001)	0.59 (.002)	0.75 (.001)	98.41 (.027)
0.8	0.97 (.003)	0.71 (.005)	0.79 (.002)	98.03 (.051)	0.98 (.001)	0.72 (.002)	0.81 (.001)	98.47 (.033)
1.0	0.98 (.001)	0.79 (.004)	0.84 (.002)	97.91 (.060)	0.99 (.001)	0.79 (.002)	0.85 (.001)	98.31 (.036)
1.2	0.99 (.001)	0.88 (.002)	0.88 (.001)	97.81 (.056)	0.99 (.000)	0.88 (.002)	0.89 (.001)	98.15 (.040)
1.4	0.99 (.000)	0.91 (.002)	0.91 (.001)	97.96 (.043)	1.00 (.000)	0.90 (.002)	0.91 (.000)	98.43 (.031)
1.6	0.99 (.000)	0.90 (.002)	0.93 (.001)	98.27 (.043)	1.00 (.000)	0.89 (.002)	0.93 (.000)	98.53 (.028)

Note. All statistics were computed using final estimated proficiencies (θ).

5.2 Research Question 2

The second research question demonstrates the practical utility of the proposed conditional statistics for the amount of adaptation as diagnostic tools for improving the adaptivity of a CAT. To do this, a hypothetical scenario of a state-wide testing program was introduced in Section 4.3. It would be expected that findings from this demonstration can inform us of some insights about how many items need to be added to achieve an acceptable level of adaptation in a particular proficiency region of interest.

5.2.1 Baseline for the CATs

As a first step, results for the current 40-item CATs administered to 2,000 examinees using the 300-item pool were evaluated in terms of three perspectives of conditional adaptivity and measurement precision. The statistics were computed to be served as a baseline to determine the proficiency levels where the amount of adaptation was not adequate during the CAT administration. Figure 5.17 presents the distributions of three conditional adaptation measures over the proficiency continuum. Based on the benchmark values for the 3PL model with MLE, two proficiency regions, colored by red in the plot, were selected where either DOD, CPRV, or ROI was below the suggested criteria for good adaptation: (1) $-0.25 <$ proficiency (θ) < 0.25 and (2) $1.75 <$ proficiency (θ) < 2.25. Out of 2,000 students, the former includes about 400 students, and the latter includes about 60 students.

For the first proficiency region ($-0.25 < \theta < 0.25$), the DOD values were below the criterion value of mid .70s, while CRPV and ROI were acceptable. It means that students in that region received items whose characteristics on average deviated to some extent from the proficiency estimate at which those items were selected, relative to the average distance between all the eligible pool items from that current estimate. As variation in the characteristics of the

administered items was small, it is plausible that the item pool did not include items whose information peaks at that proficiency region. For the other region ($1.75 < \theta < 2.25$), although the DOD and ROI values were acceptable, the CPRV value was lower than the criterion value of low 0.80s. It indicates that while the students took items whose characteristics were generally well matched, the item pool did not include sufficiently good items for the students in that region so that the item selection algorithm may have to select some items whose characteristics poorly match their interim proficiency estimates. In addition, given the bias, TSEM, and RMSE values (see Figure 5.18), the measurement accuracy and precision of the proficiency estimates in these two regions were similar. Therefore, it would be interesting to see how many items need to be added in the item pool to approach to the acceptable levels of adaptivity in these two regions.

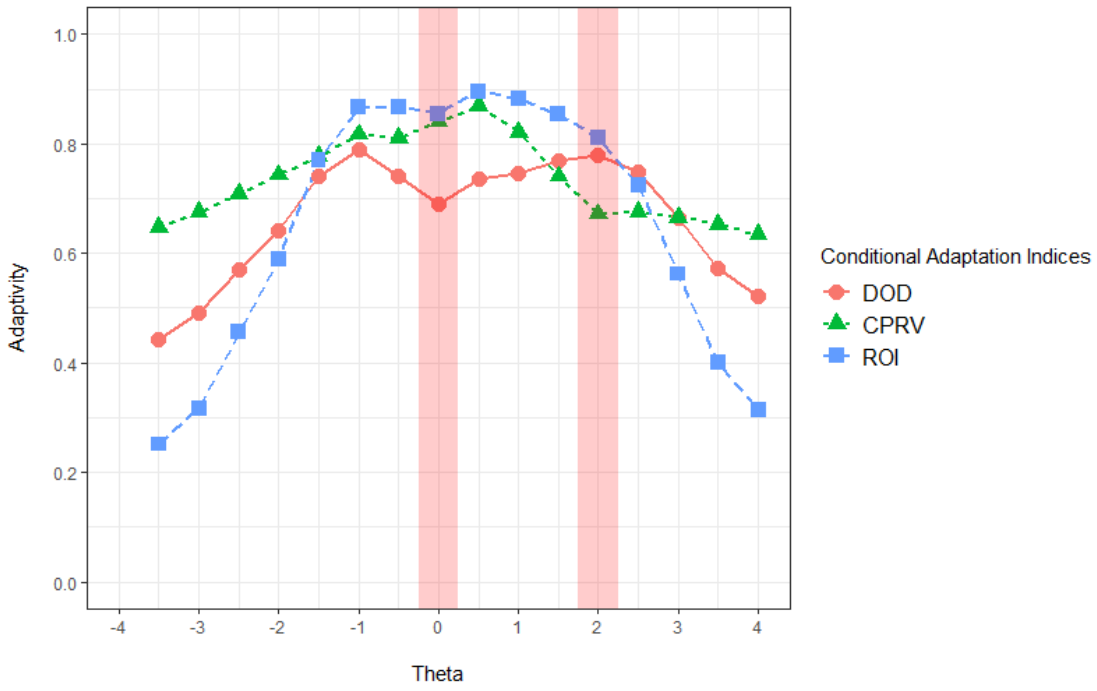


Figure 5.17. A plot of conditional adaptivity indices over the proficiency continuum for the CAT using the 300-item pool (baseline).

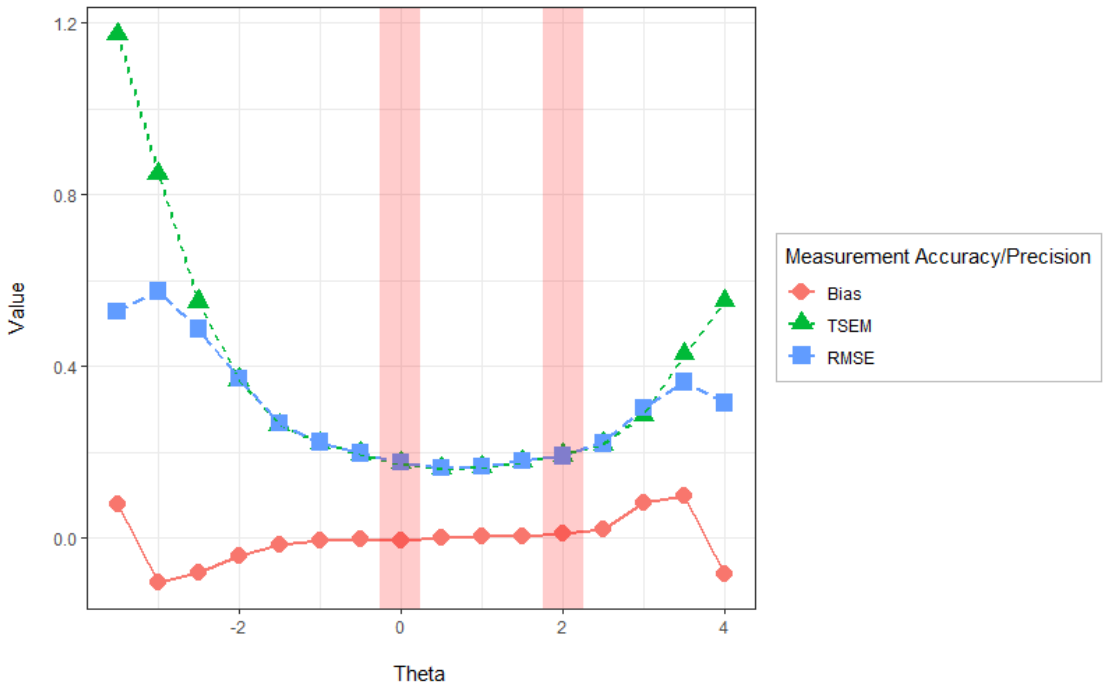


Figure 5.18. A plot of bias, TSEM, and RMSE over the proficiency continuum for the CAT using the 300-item pool (baseline).

Once the proficiency region of interest was determined, a series of fixed numbers of items were added to attain an acceptable level of adaptation at each region. The items to be added were selected whose information was high at that proficiency region from the master pool. Note that to mimic the real-word situation, the items in the master pool were normally distributed, instead of uniform information. Hence, the quality of items to be added was not fully controlled so that the item quality may not be equal across the items to be selected from the master pool. For each region, the fixed numbers of items to be added to the item pool are 5, 10, 15, 20, 30, 40, 50, and 100, and the results were replicated over 50 times.

5.2.2 Region 1: -0.25 < θ < 0.25

As a starting point, the mean value of each adaptivity index at the first proficiency region (-0.25 < θ < 0.25) was 0.69 (*SD* = .01) for DOD, 0.84 (*SD* = .02) for CRPV, and 0.85 (*SD* = .01)

for ROI. While the CPRV and ROI were acceptable, the DOD index was below the criterion of mid 0.70s suggested in the study for Research Question 1 (see Section 5.1). To improve adaptivity to the acceptable levels, 5, 10, 15, 20, 30, 40, 50, and 100 items from the master pool were sequentially added to the existing operational item pool. Results reported that compared to the baseline for the three conditional adaptation statistics, as the items that were the most informative in the master pool at that proficiency region were added to the operation item pool, the three aspects of adaptivity were gradually improved (see Figure 5.19). In particular, when 30 informative items were additionally included to the operational pool, the DOD values at the Region 1 of students showed clear, visible improvement, exceeding the benchmark value of mid .70s, and the other two statistics also obviously increased. The three statistics were not much improved after more than 40 items were added to the item pool.

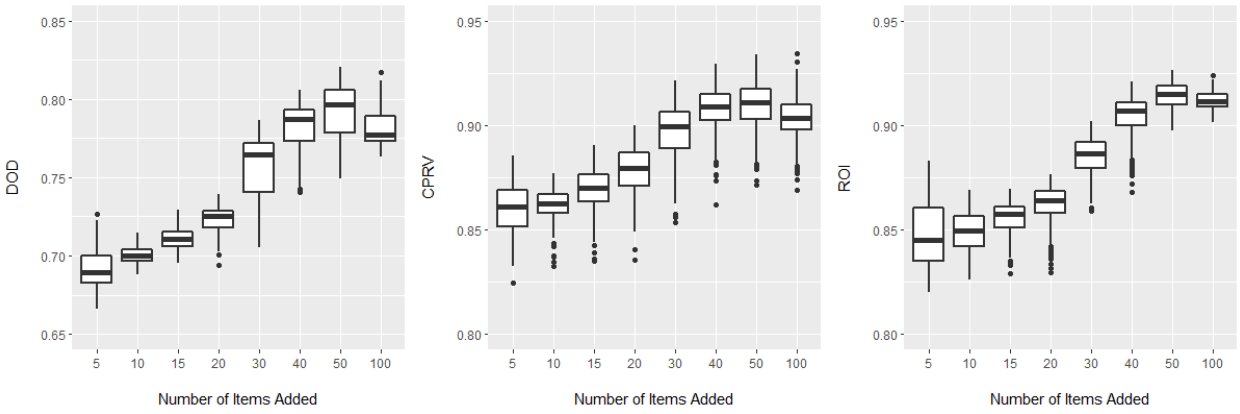


Figure 5.19. Distributions of conditional adaptivity indices by number of items added at Region 1 ($-0.25 < \theta < 0.25$).

Regarding the measurement accuracy and precision, a similar pattern was identified. As more items were added to the item pool, the bias and standard errors decreased. In particular, similar to the distributions of the adaptivity indices, when 30 items were additionally included to the operational pool, the standard errors (TSEM, RMSE) were reduced. This suggests that the

improvement of adaptivity for the CAT can lead to enhancing the measurement precision of proficiency estimates.

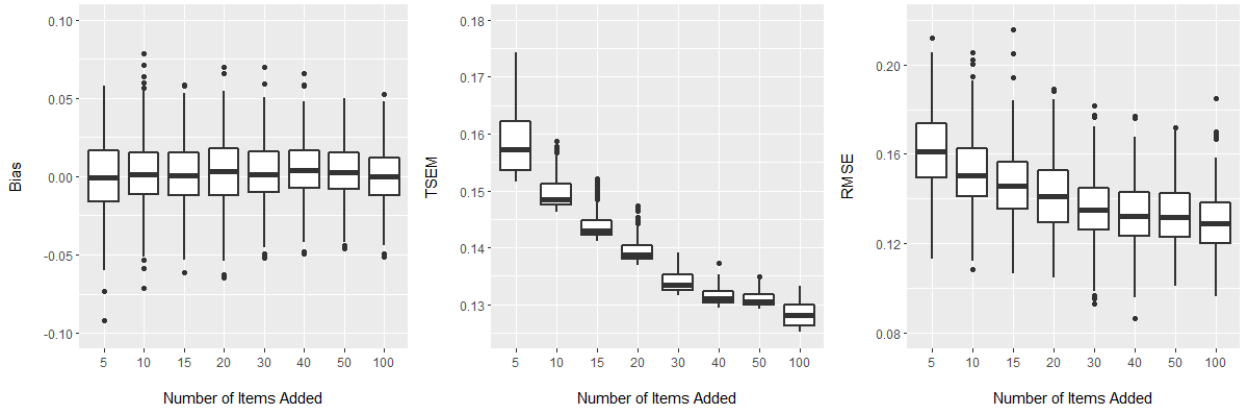


Figure 5.20. Distributions of statistics for measurement accuracy and precision by number of items added at Region 1 ($-0.25 < \theta < 0.25$).

5.2.3 Region 2: $1.75 < \theta < 2.25$

Another proficiency region that need to be improved is the proficiency levels ranging from 1.75 to 2.25, called Region 2. For the baseline at Region 2 ($1.75 < \theta < 2.25$), the mean value of each adaptivity index was 0.78 (*SD* = .004) for DOD, 0.67 (*SD* = .007) for CRPV, and 0.81 (*SD* = .012) for ROI. In contrast with the results for Region 1, the DOD values were acceptable, whereas the CPRV values were lower than the guideline of low 0.80s and the ROI was also slightly below the benchmark of mid 0.80s.

To improve all adaptivity indices to be acceptable, again, 5, 10, 15, 20, 30, 40, 50, and 100 items from the master pool were sequentially added to the existing operational item pool. As a result, compared to the baseline for the three conditional adaptation statistics, as more items that were the most informative in the master pool at that proficiency region were included in the operation item pool, the three adaptivity indices clearly increased (see Figure 5.21). In particular, when 30 items that were the most informative at Region 2 among the eligible items in the master

pool were additionally included to the operational pool, the CPRV values increased but were still below the acceptable level. After adding 50 items, the CPRV reached an acceptable level of adaptivity. As with the findings for the CPRV, the DOD and ROI measures presented visible enhancement after 30 items were additionally included in the operational item pool. The ROI index exceeded the benchmark value when at least 40 items were added to the item pool. One thing that should be noted is that the DOD value decreased when 100 items were added. Although the variances in the eligible items for the pool were similar, as the quality of items in the master pool was not fully controlled, some items added were relatively informative to those students at Region 2 but the location of these items whose information was optimal deviated from the current proficiency estimates. This might be plausible because of the impact of the *a*-parameter on the information function. It thus resulted in reducing the DOD value.

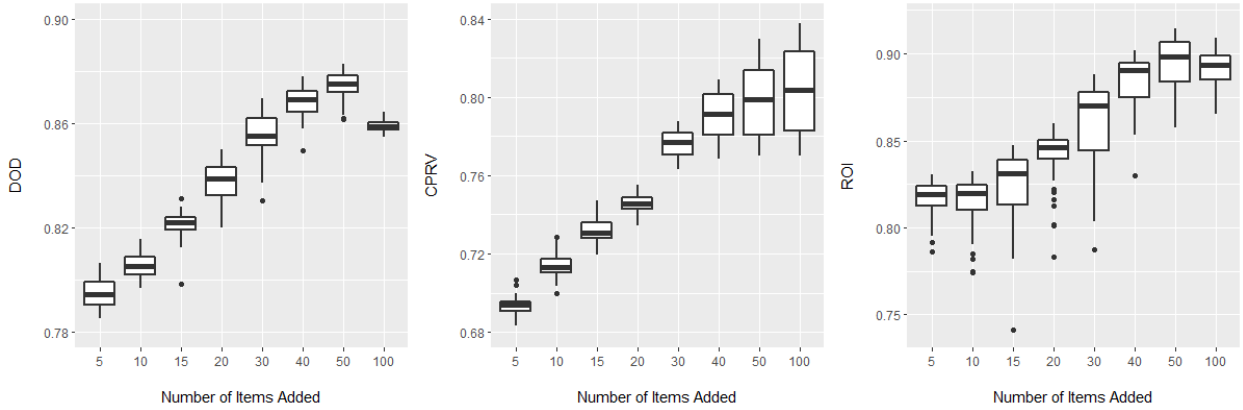


Figure 5.21. Distributions of conditional adaptivity indices by number of items added at Region 2 ($1.75 < \theta < 2.25$).

The resulting distributions of bias, TSEM, and RMSE for Region 2, shown in Figure 5.22, were consistent with the results for the Region 1 study. The more informative items there are in the operation item pool, the smaller bias and standard errors of the proficiency estimates there are. Particularly, the TSEM values were apparently reduced when 30 informative items

were added to the operational item pool. After more than 40 items were added, the measurement precision was not much improved. Comparing the adaptivity to the measurement precision, it provides some evidence showing that the amount of adaptation for CATs would not always function together with the measurement precision.

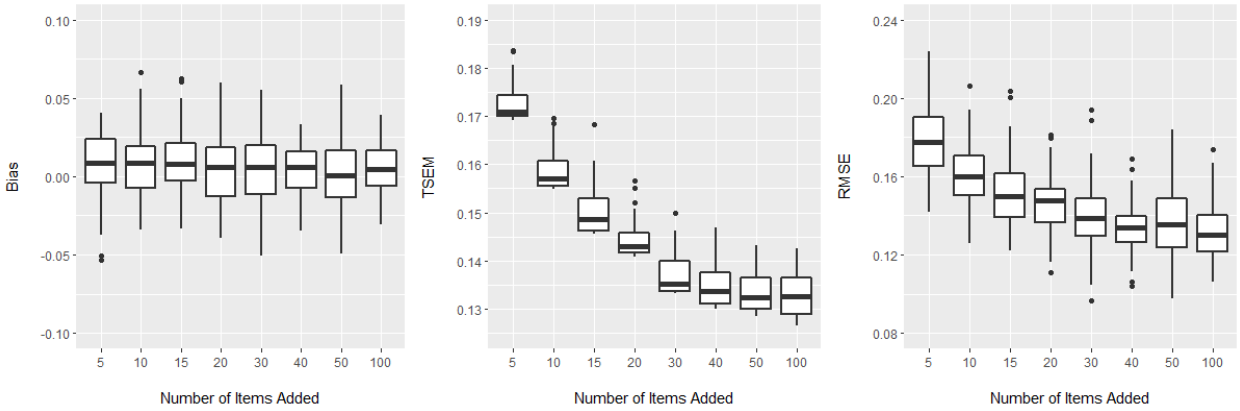


Figure 5.22. Distributions of statistics for measurement accuracy and precision by number of items added at Region 2 ($1.75 < \theta < 2.25$).

To sum up, the results for these simulation studies can answer how many items need to be added to improve the adaptivity for the CATs. Although the item quality was not fully controlled in the study, it suggests that for the given CAT specifications and the item pool distribution, in general, adding about 30 items that are informative at the particular proficiency levels of interest can contribute to visible enhancement of the amount of adaptation. However, it should be noted that the number of items can be adjusted depending on the item quality.

5.3 Research Question 3

Exposure control is a vital aspect of CAT for test security purpose in large-scale assessments. Because of the limited availability of computers, test sessions are usually scheduled multiple times per day or every day over the week. That means that examinees can share information about the test items that they have taken before and after their tests, resulting in threats to test security. To tackle this concern, exposure control procedures have been suggested as a way of putting some constraints on item selection to limit the number of items that students can share in common. Such constraints might affect the amount of adaptation of a CAT by preventing it from selecting the best items that match well with the current proficiency estimate.

This chapter primarily explores the effect of the exposure control on the level of adaptivity for a CAT and then further investigates whether these effects can be moderated by the item pool characteristics. Three exposure control procedures that are commonly used in practice were considered in this study: (a) the randomesque procedure, (2) *a*-stratified method with *b*-blocking (BAS), and (3) the Sympson-Hetter method. A CAT with no exposure control procedure was administered for comparison purposes. Also, two 300-item pools with different shapes of distributions of item difficulty were employed: (1) a bell-shaped regular item pool, and (2) a rectangular-shaped optimal item pool that was created using the bin-and-union method (see Section 4.4.1.1 for technical details). The results were summarized in terms of the relation between true and estimated proficiency (i.e., measurement accuracy/precision), the amount of adaptation, and test security in the following sections.

5.3.1 Measurement accuracy and precision

Conditional statistics. Figure 5.23 displays the measurement accuracy and precision of proficiency estimates contingent on proficiency level in term of bias, TSEM, and RMSE. Results

indicated that regardless of exposure control procedures, there was little bias in proficiency estimates over the proficiency scale except for the extreme ends of the scale. Compared to the CAT performance with no exposure control, the BAS and Sympson-Hetter methods produced slightly larger standard errors of the proficiency estimates than the randomesque procedure did, which was consistent when either item pool was used. However, with the well-designed optimal item pool, the standard errors were noticeably reduced especially at the extreme proficiency regions, suggesting the proficiency estimates were nearly equally precise across the proficiency levels.

Overall statistics. Without exposure control, the CAT using the optimal item pool reported slightly smaller standard errors (TSEM and RMSE), implying that proficiency was more precisely estimated over the entire sample of examinees. Compared to the results of CAT with no exposure control, all three exposure control procedures presented larger standard errors, implying less precise proficiency estimates. With the regular item pool, the BAS design provided larger RMSE and smaller fidelity ($r_{\theta\hat{\theta}}$), whereas with the optimal item pool the Sympson-Hetter approach reported larger RMSE and smaller fidelity. Unlike the other two exposure control procedures, the Sympson-Hetter method did not result in a difference in overall measurement precision between the regular item pool and the optimal item pool.

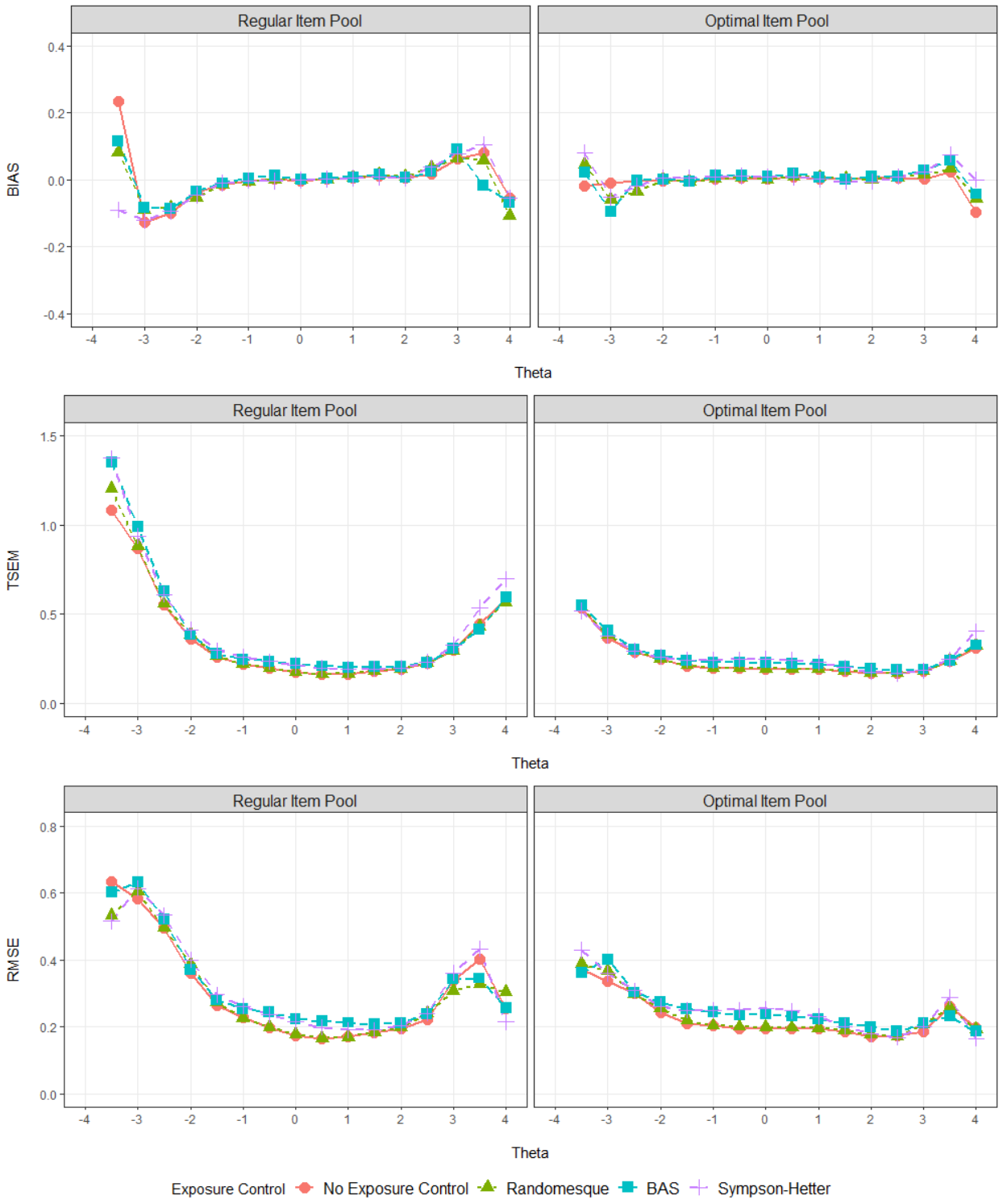


Figure 5.23. Conditional bias, TSEM, and RMSE of proficiency estimates for the 3PL-based 40-item CAT by exposure control procedure and item pool distribution.

Table 5.9

Overall Statistics of Measurement Precision of Proficiency Estimates for the 3PL-based 40-item CAT by Exposure Control Procedure and Item Pool Distribution

Item Pool	Exposure Control Procedure	Statistic			
		Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
Regular item pool	No exposure control	-0.002	0.209	0.207	0.980
	Randomesque procedure	-0.002	0.213	0.211	0.979
	a-stratification with *b*-blocking	0.004	0.245	0.244	0.972
	Sympson-Hetter method	-0.002	0.245	0.242	0.973
Optimal Item pool	No exposure control	0.003	0.197	0.199	0.981
	Randomesque procedure	0.003	0.200	0.203	0.980
	a-stratification with *b*-blocking	0.010	0.228	0.236	0.974
	Sympson-Hetter method	0.006	0.240	0.244	0.972

5.3.2 Amount of adaptation

Conditional adaptivity. As shown in Figure 5.24, in general, using the well-designed optimal item pool, higher adaptability was attained over a broad range of the proficiency continuum across the exposure control procedure conditions. Regardless of item pool characteristics, the BAS design for exposure control led to an improvement in adaptivity over the proficiency continuum even compared to the CAT procedure with no exposure control. The DOD and ROI indices noticeably increased over the proficiency levels ranging from -1.0 to 2.0 for the regular item pool and over the proficiency ranging from -3 to 3 for the optimal item pool. This is expected to some degree in that for the BAS, the items were selected whose *b*-parameters had a good match with the current proficiency estimate, which is closely associated with the concepts of the DOD index. Also, the BAS forced items with high *a*-parameters providing more information than those with the low *a*s, to be used in later stages of a test. Since the efficiency of an item with high *a* might not be fully utilized if the (true) proficiency is not close to the

difficulty of that item (Hambleton & Swaminathan, 1985, pp. 108-115), the item with high *a*-parameter should be used later in the test when more accurate proficiency estimates are available. Accordingly, the BAS yielded higher ROI values, suggesting that a test was efficiently adapted to each examinee.

However, the Sympson-Hetter method reduced the level of adaptation regardless of item pool characteristics. The DOD values were dramatically low. A plausible explanation is that the Sympson-Hetter method controls for overexposed items by distinguishing the item selection and administration processes, but it does not contribute to using underexposed or never used items that are rarely selected based on the maximum information item selection criterion. For the 3PL model an item with high *a*- and small *c*-parameters usually has high information, resulting in that item being more likely selected and administered even though the *b*-parameter (or location of the maximum) of that item is not closely matched with the current proficiency level. This poor usage of informative items (i.e., mismatch between the item location and the current proficiency estimate) can be made even worse by limiting the highly exposed items by the Sympson-Hetter procedure so that the DOD values yielded were very low. The randomesque procedure did not have much effect on the level of adaptation over the proficiency continuum.

Overall adaptivity. The overall adaptivity results are summarized in Table 5.10. With no exposure control, the CAT procedure clearly performed very well with the designed optimal item pool rather than with the regular item pool. The latter reported the ratio of SDs index, 0.74, slightly below the suggested benchmark value of high 0.70s. With the optimal item pool, all the overall adaptivity measures were very high even though the exposure control was imposed on the item selection for the CAT. Regardless of item pool characteristics, the BAS led to the increase in the ratio of SDs index but with a small decrease in the correlation and PRV indices. The

former may be due to the fact that items were selected from only one quarter of the full item pool across stages of a test. The latter may be because selecting items within each stratum leads to the selection of more extreme items as *b*-blocking makes extreme items available for the item selection. In addition, the Sympson-Hetter method reduced the PRV values. As previously mentioned, the Sympson-Hetter method limits overexposed items to keep the exposure under the pre-specified value, leading to items being administered to each examinee on average that are more widely spread around their final proficiency estimates than the other procedures. However, the PRV value was still above the criterion of good adaptation with the optimal item pool, suggesting the sensitivity of the Sympson-Hetter method to the item pool characteristics. Again, as with the conditional adaptivity measures, it was observed that the randomesque procedure barely affected the amount of adaptation for a CAT for the entire group level of examinees.

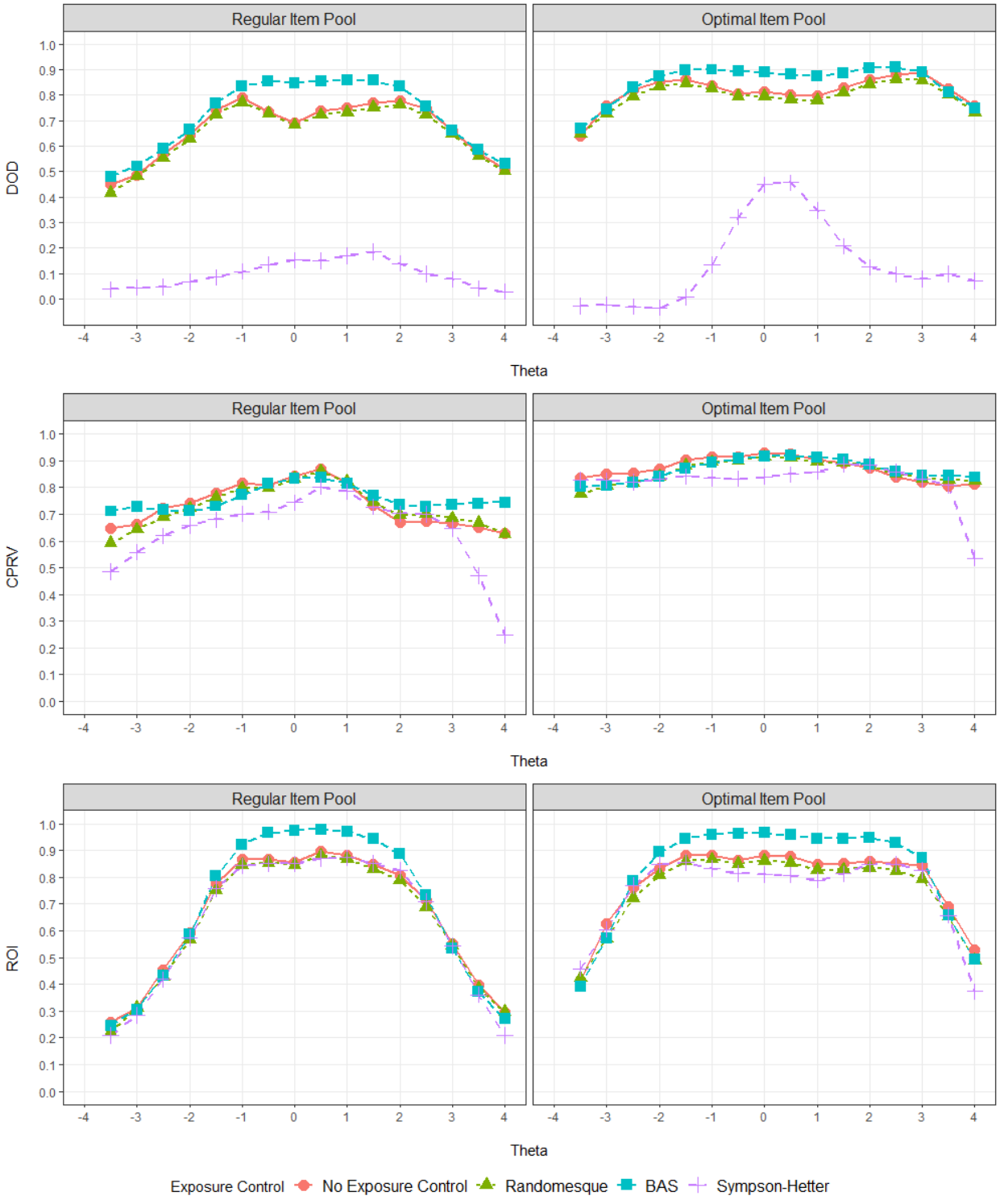


Figure 5.24. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a 3PL-based 40-item CAT by exposure control procedure and item pool distribution.

Table 5.10

Overall Adaptation Statistics for a 3PL-based 40-item CAT by Exposure Control Procedure and Item Pool Distribution

Item Pool	Exposure Control Procedure	Statistic		
		$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
Regular item pool	No exposure control	0.97 (.002)	0.74 (.004)	0.81 (.002)
	Randomesque procedure	0.97 (.002)	0.73 (.005)	0.80 (.002)
	a-stratification with *b*-blocking	0.94 (.002)	0.80 (.006)	0.80 (.003)
	Sympson-Hetter method	0.96 (.004)	0.74 (.005)	0.74 (.002)
Optimal Item pool	No exposure control	0.99 (.001)	0.91 (.003)	0.91 (.001)
	Randomesque procedure	0.99 (.001)	0.91 (.003)	0.90 (.001)
	a-stratification with *b*-blocking	0.95 (.002)	0.95 (.005)	0.90 (.001)
	Sympson-Hetter method	0.98 (.002)	0.99 (.004)	0.85 (.001)

5.3.3 Test security

As shown in Figure 5.25, the Sympson-Hetter method presented better exposure control of the highly exposed items than the other procedures but did not successfully control for the underexposed or unused items. However, the BAS approach had well-balanced item exposure and better utilization of items, showing a decrease in exposure rates for the items that were highly exposed and an increase in exposure rates for the items that were rarely or never used compared to the results with no exposure control. It is noted that the BAS approach had more items whose exposure rates were greater than 0.20 for the optimal item pool rather than for the regular item pool. The randomesque procedure led to reducing the overexposure rates of the items whose difficulty was around 0.0. Without the exposure control procedure, all examines had the same initial proficiency estimates of 0.0, resulting in taking the same first item with *b*-parameter closest to 0.0. That first item reported the perfect exposure rate shown in Figure 5.25.

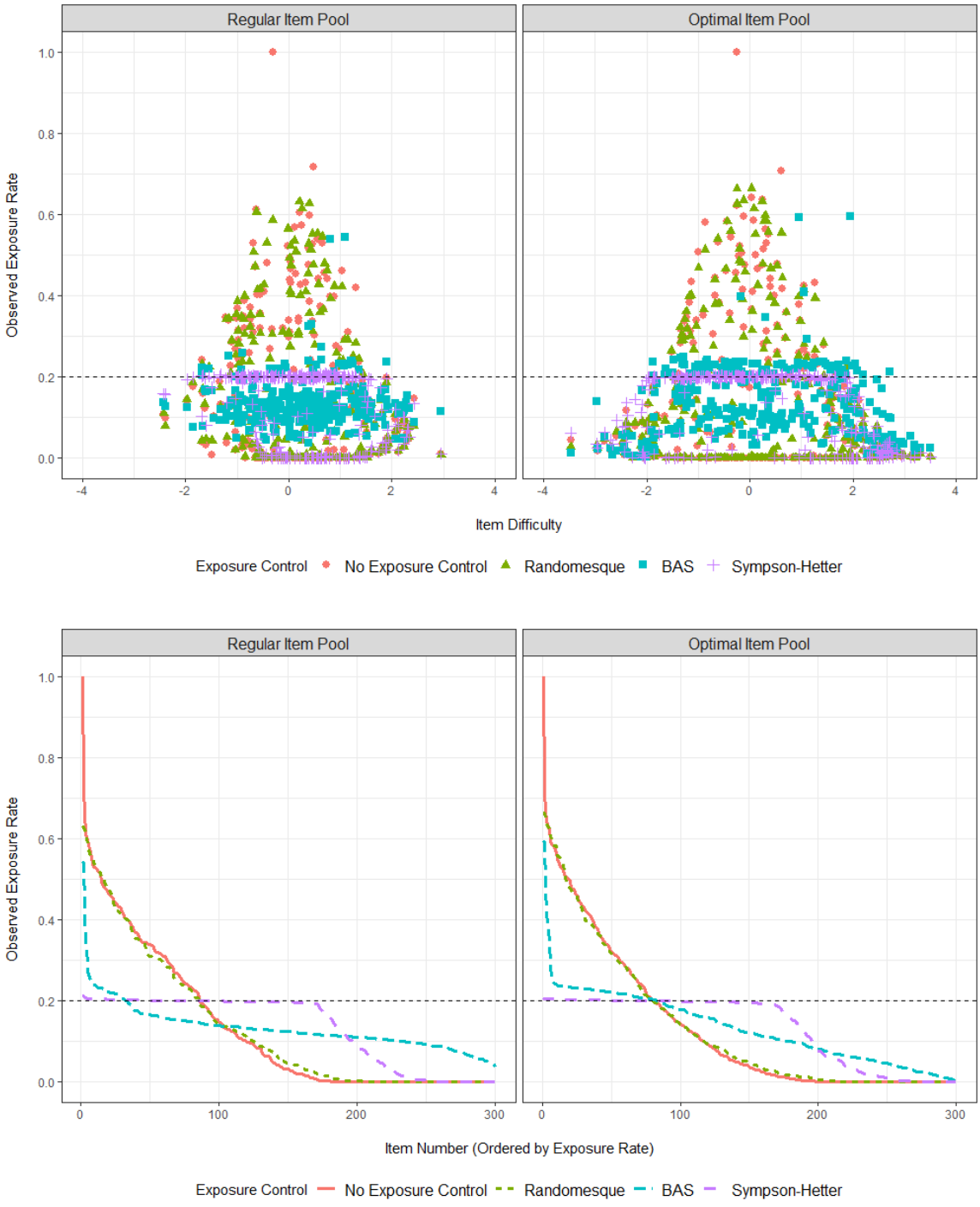


Figure 5.25. Exposure rate distribution of 300 items ordered by *b*-parameter (top) and exposure rate (bottom) for a 3PL-based 40-item CAT by exposure control procedure and item pool distribution.

5.4 Research Question 4

To further demonstrate the utility of the proposed measures of conditional adaptivity with the benchmark values obtained in the first research question, this section for Research Question 4 evaluates the functioning of different adaptive test designs moderated by item pool characteristics in terms of the property of proficiency estimates and the amount of adaptation. In this study, two adaptive testing designs were considered: (1) an item-level CAT, in which individual items are fully adapted to an examinee's proficiency estimate during the CAT procedure; and (2) a multistage adaptive test (MST), which is adapted to the stage or module (i.e., a set of items) level of items for an examinee. As previously mentioned, the 1-2-3 three-stage MST with increasing module length through stages (i.e., 10-10-20 items) was constructed by selecting 90 items from an item pool. It was hypothesized that MST reported less adaptivity than item-level CAT, although MST can achieve a similar level of measurement precision. In addition, as with the study for Research Question 3, the same two item pools (i.e., regular item pool and optimal item pool) were used.

5.4.1 Measurement accuracy and precision

Conditional statistics. Figure 5.26 presents the distributions of the bias, TSEM, RMSE values across the proficiency continuum by test design and item pool characteristics. In general, both test designs had comparable measurement accuracy and precision in the moderate proficiency range; however, the MST design obviously had more bias and standard errors at the extreme regions of the proficiency scale than the fully adaptive CAT. With the optimal item pool, the CAT reported smaller values of all three statistics than those with the regular item pool and had even measurement precision (i.e., TSEM) over the entire proficiency continuum. Meanwhile, the MST did not make big differences in the recovery of proficiency estimates

between the two item pools, reporting fairly high standard errors and bias at the extreme ends of the proficiency scale.

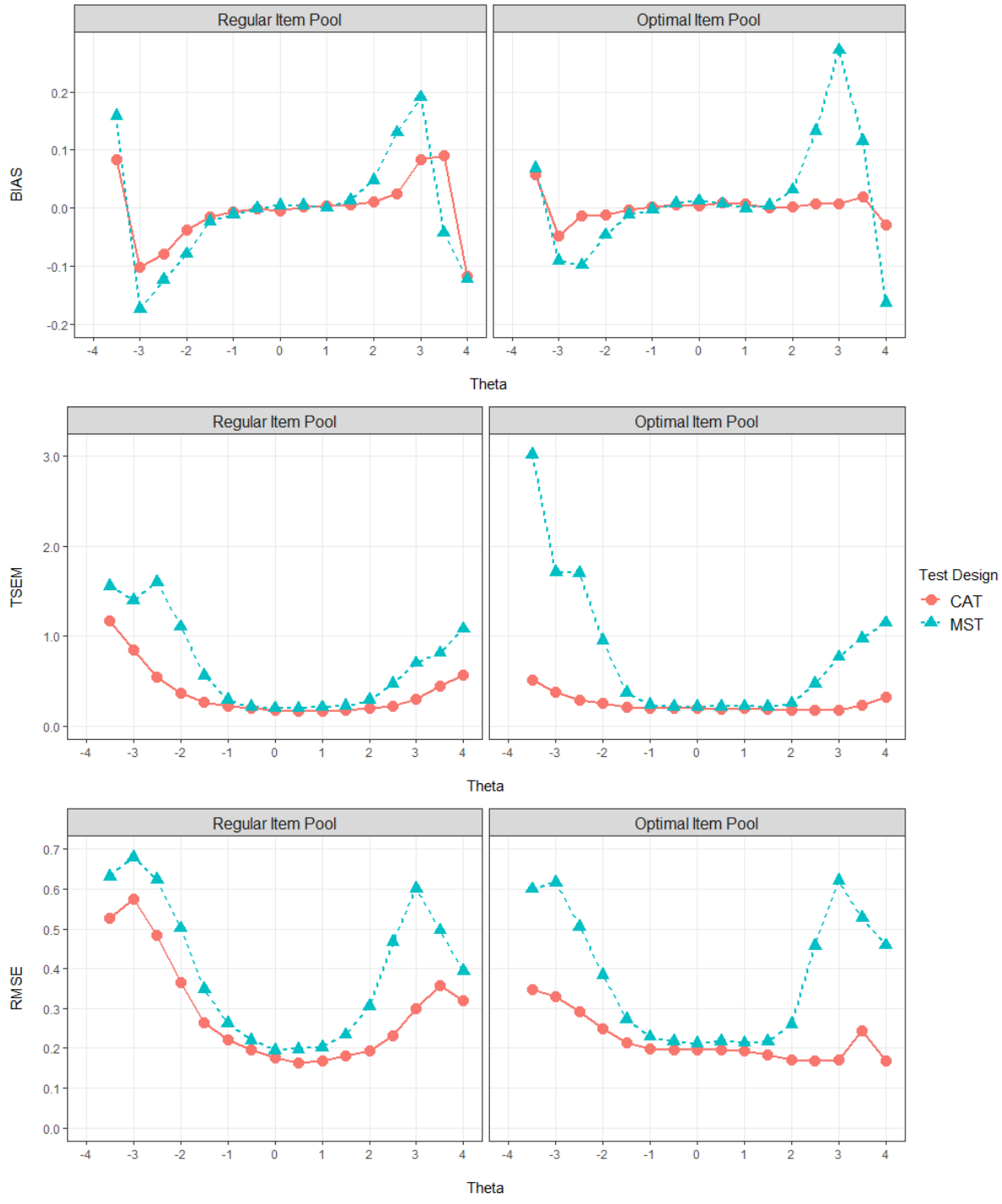


Figure 5.26. Conditional bias, TSEM, and RMSE of proficiency estimates for a 3PL-based adaptive test by test design and item pool distribution.

Overall statistics. Results indicated that regardless of item pool characteristics, the fully adaptive testing (CAT) reported better recovery of proficiency estimates giving lower TSEM and RMSE values and higher fidelity correlation than the MST. For the item-level CAT, the optimal item pool contributed to slightly improved measurement precision of proficiency estimates relative to the typical operational item pool. Although the MST design created from the regular item pool showed a similar pattern of conditional measurement statistics to the MST using the optimal item pool, the overall statistics, except for the bias, indicated slightly better measurement precision for the latter.

Table 5.11

Overall Statistics of Measurement Precision of Proficiency Estimates for the 3PL-based 40-Item Adaptive Test by Test Design and Item Pool Distribution

Item Pool	Test Design	Bias	TSEM	RMSE	$r_{\theta\hat{\theta}}$
Regular Item Pool	Full CAT	-0.003	0.209	0.207	0.980
	MST	-0.001	0.420	0.259	0.970
Optimal Item Pool	Full CAT	0.003	0.197	0.200	0.981
	MST	0.005	0.402	0.243	0.973

5.4.2 Amount of adaptation

Conditional adaptivity. Which examinees were not administered items of appropriate quality? Putting it differently, how can we improve the test designs or the quality of an item pool for better adaptivity across the proficiency levels of interest? The conditional adaptation measures proposed in this study can help tackle these concerns. As shown in Figure 5.27, the CAT gave items better adapted for students' proficiency levels than the MST from three aspects of adaptivity across proficiency levels, though they had similar level of adaptation at some

proficiency regions. In fact, the MSTs did not satisfactorily meet the guidelines of the DOD (i.e., mid .70s), CPRV (low .80s), and ROI (mid .80s) measures, and their adaptivity seemed different across the individual proficiency levels. For instance, adaptivity around the proficiency of $\theta = 0$ was better than other proficiency regions in that informative items were properly administered and adapted for the students according to the CPRV and ROI values close to the guidelines. On the contrary, the CAT using the optimal item pool yielded equally high values of the proposed three indices that exceed the guidelines over the broader range of the proficiency levels than the results of the CAT using the regular item pool. The latter met the suggested benchmark values of good adaptation in the moderate proficiency range but did not in other proficiency areas.

Overall adaptivity. Table 5.12 summarizes the overall measures of adaptation by test designs and item pools. As with the results of conditional adaptivity, using either item pool, the values of correlation, ratio of SDs, and PRV indices for the MST design were notably lower than those for CAT. This is because MSTs adapted at the module/stage level while CATs are adapted at the item level. Also, all students took the same routing module of 10 items, and students who took the same path through the stages received identical test items, which can contribute to limiting the amount of adaptation to some extent. For the item-level CAT, using the optimal item pool led to obviously improved values of the ratio of SDs and the PRV index, implying that CATs presented items well-customized to the final proficiency estimates. This is due to the fact that the optimal pool included more informative items over the entire range of examinees' proficiency, meaning a larger SD of the difficulty parameters. Interestingly, for the MST, the value of ratio of SDs for the optimal pool was greater than that for the regular pool due to the same reason previously mentioned, whereas the other two indices were comparable to one another. This implies that the optimal pool allowed students to take on average the items that

more closely matched to their proficiency level but showed similar degree of other aspects of adaptivity to the regular item pool. It should be noted that overall adaptation indices for either MST were apparently lower than the guidelines, implying not as good as adaptation of CAT.

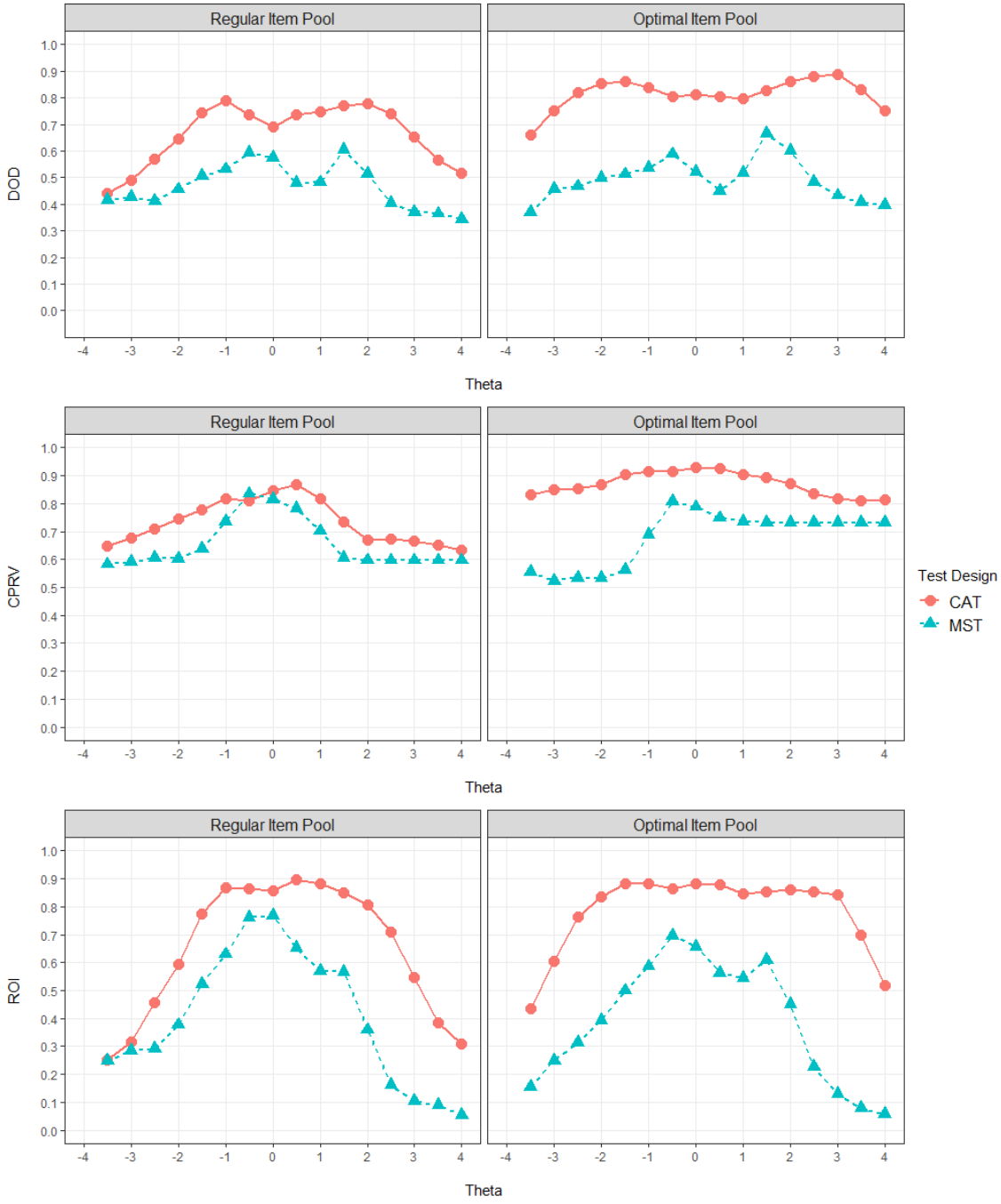


Figure 5.27. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a 3PL-based 40-item adaptive test by exposure control procedure and item pool distribution.

Table 5.12

Overall Adaptation Statistics for a 3PL-based 40-item CAT by Exposure Control Procedure and Item Pool Distribution

Item Pool	Test Design	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
		M (*SE*)	*M* (*SE*)	*M* (*SE*)
Regular Item Pool	Full CAT	0.97 (0.002)	0.74 (0.004)	0.81 (0.002)
	MST	0.88 (0.005)	0.55 (0.008)	0.75 (0.002)
Optimal Item Pool	Full CAT	0.99 (0.001)	0.91 (0.003)	0.91 (0.001)
	MST	0.88 (0.006)	0.65 (0.007)	0.74 (0.001)

5.5 Research Question 5

Lastly, this section demonstrates whether both conditional and overall adaptivity statistics function as expected using the real operational dataset to examine the amount of adaptation over the entire sample of examinees and at the individual proficiency levels. To do this, the finial proficiency estimates of examinees, the list of item parameters for the items administered to each examinee, the list of item scores or interim proficiency estimates for the items administered to each examinee, and the item parameters for the item pool are required for computing conditional and overall statistics. These data were available for the NCLEX-RN licensure examination in 2017, which employed a variable-length CAT with a minimum test length of 60 and a maximum length of 250 operational items. Note that the pretest items were removed from this analysis. The total sample for this administration period included about 70,000 examinees. Like a previous study (Reckase et al., 2018), 35 subsamples of 2,000 examinees were randomly sampled from the full sample without replacement. This allowed me to evaluate the stability of the adaptation measures, as well.

5.5.1 Conditional adaptivity

The conditional adaptivity indices clearly provided evidence about how the NCLEX exam was designed for their classification (pass/fail) purpose based on the cut-off scores of 0.0. Around the cut-score of $\theta = 0$, all three adaptivity measures met the guidelines suggested in the study for Research Question 1, implying that the test provided informative items efficiently customized to classify students whose proficiency was above or below their criterion, 0.0. To be specific, the DOD value was slightly below the guideline of mid .90s, but the CPRV and ROI measures were far better than the guidelines at the cut-off value of $\theta = 0.0$. In addition, these measures were stable across 35 samples of 2,000 examinees randomly sampled from the full dataset. Their empirical standard deviations were in the ranges of 0.04 to 0.12 for the CPRV index, 0.02 to 0.06 for the DOD index, and 0.01 to 0.07 for the ROI index. Compared to the other two statistics, the CRPV showed slightly more variations in the measure across samples, which might due to the property of the proficiency estimation procedure used in the NCLEX-RN test. That is, the Owen Bayesian estimation (Vale & Weiss, 1977) with a prior with a mean of -1.0 and a standard deviation of 2.0 was used in the beginning of the test and then the MLE was employed after both correct and incorrect responses exist for an examinee. In this procedure, the selected items were affected by the current proficiency estimate, causing more within-examinee variation in *b*-parameters of the CPRV. At any rate, there was no doubt that the NCLEX-RN test reported outstanding adaptation at their cut point, which is well aligned with their test purpose.

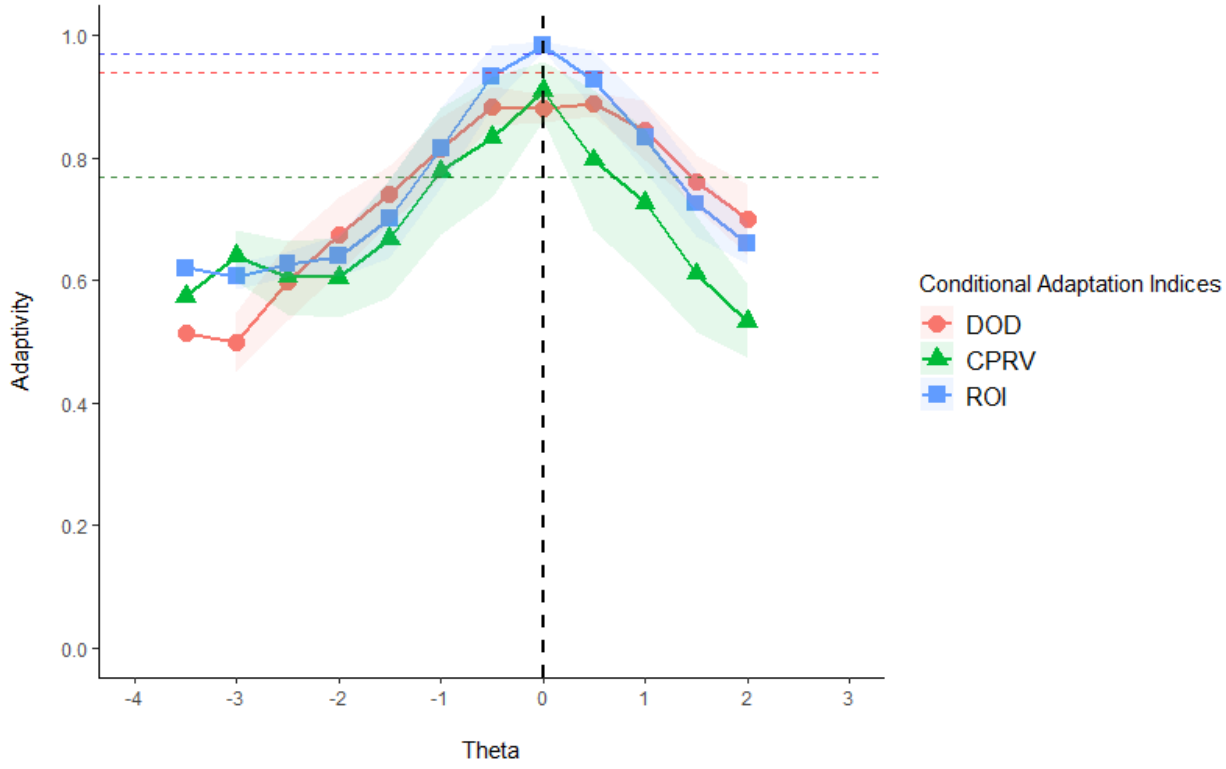


Figure 5.28. Conditional adaptivity statistics (DOD, CPRV, and ROI) for a Rasch-based variable-length CAT for an operational NCLEX-RN test.

5.5.2 Overall adaptivity

Table 5.13 summarizes the results for the overall values to gauge the level of adaptation for the NCLEX test over the sample of examinees during this administration period. For three overall adaption measures, this test complied with the guidelines suggested by the simulation studies for Research Question 1 in Section 5.1. Specifically, the correlation index was 0.91, which was in the low 0.90s, the ratio of SDs index was 0.90, which exceeded the benchmark value in the mid 0.80s, and the PRV value met the guideline of about 0.80. Taken all together, these indices indicated that this NCLEX test is worthy of being labeled an “adaptive” test. Despite the constraints of the item selection algorithm for content balancing and exposure

control, this test showed good adaptivity resulting from the well-designed item pool and a strong item selection algorithm.

Table 5.13

Overall Adaptation Statistics for a Rasch-Based Variable-Length CAT for an Operational NCLEX-RN Test

	$r(\bar{b}_j, \hat{\theta}_j)$		$s_{\bar{b}_j}/s_{\hat{\theta}_j}$		PRV	
	M	SD	M	SD	M	SD
NCLEX	0.91	0.004	0.90	0.010	0.80	0.003
Benchmark Values	Low 0.90s		Mid 0.80s		0.80	

Note. SD = Empirical standard deviation over 35 samples of data.

CHAPTER 6.

CONCLUSION AND DISCUSSION

6.1 Summary of Findings

This study's aim was twofold. First, it was to propose new statistical indicators to measure the amount of adaptation conditional on proficiency levels for computerized adaptive testing (CAT). Second, it was to evaluate their feasibility and utility at detecting how much a test is customized to a student's proficiency. This customization is a function of the quality of item pool, proficiency estimators, constraints on item selection, and test design through simulations and empirical demonstration using real data analysis. Three conditional adaptation measures were (1) the deviation of difficulty (DOD), (2) the conditional proportion of reduction in variance (CPRV) index, and (3) the ratio of information (ROI). The proposed measures provide slightly different information about the amount of adaptation assessed over the entire sample of examinees using the existing overall adaptation indices. These measures can help us understand the adaptivity of a test for an individual examinee or for subgroups of particularly interest. For a particular subgroup of students, for example, the CAT reported a high CPRV value but low DOD and ROI values. In these test events, it is plausible that the students received items of similar difficulty, but the administered items deviated, on average, widely from their provisional proficiency estimate. This might be because an item pool did not contain informative items for those students or there were some problems with the item-selection procedure. Taken together, these conditional statistics function toward the goal of gauging the differences in the amount of adaptation from these three viewpoints. From both the simulation studies and real data analysis, five key findings were drawn.

First, the results of comprehensive simulations in Research Question 1 suggest some guidelines for interpreting the proposed conditional adaptation indices for adaptive testing by different IRT models and proficiency estimators. When the Rasch model is used for scaling and scoring, these benchmark values, as summarized in Table 6.1, indicate good adaptation—a DOD value in the mid 0.90s, with a maximum likelihood estimation (MLE) for proficiency estimator, a CPRV value in the high 0.70s, and a ROI value in the high 0.90s. Not only that, but a DOD in the mid 0.70s, a CPRV in the low .80s, and a ROI in the mid 0.80s indicate good adaptation when the 3PL IRT model is used, meaning that an adaptive test administers items that are well customized to an individual student. With the expected a posteriori (EAP) estimation method, the guidelines for the CPRV index were slightly higher than those for the MLE method, but the guidelines for the DOD and ROI indices were the same as the ones for the MLE.

Table 6.1

Benchmark Values of Conditional and Overall Adaptivity Indices by IRT Models and Proficiency Estimators

	Rasch		3PL	
	MLE	EAP	MLE	EAP
Conditional Indices				
DOD	Mid 0.90s	=	Mid 0.70s	=
CPRV	High 0.70s	High 0.80s	Low 0.80s	Mid 0.80s
ROI	High 0.90s	=	Mid 0.80s	=
Overall Indices				
$r(\bar{b}_j, \hat{\theta}_j)$	Low 0.90s	=	High 0.90s	=
$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	Mid 0.80s	=	High 0.70s	=
PRV	0.80	High 0.80s	Low 0.80s	Mid 0.80s

One thing to note is for the simulated 40-item test, these conditional statistical descriptors were stable in the middle proficiency, ranging from -2.0 to 2.0 based on the small values of empirical standard errors of the statistics. With the 3PL model, however, the statistics generally had fairly large standard errors at the extremes of the proficiency range, particularly with a small item pool or a restricted item pool. As the pool size increased and more items were located at the extremes, the stability of these measures improved but was still large compared to the error in the middle proficiency range. This instability at the extremes may be due partly to the effect of *c*-parameters in the proficiency estimation. However, this concern was resolved with the Rasch model.

The findings also provide evidence that regardless of IRT models and proficiency estimators, the bigger the item pool size is, the more spread of difficulty in the item pool there is, the better adaptation there is, and the more accurate proficiency estimates there are, which is consistent with the results of previous studies (Reckase et al., 2018; Ju & Lee, 2018). For a high level of adaptivity in CAT, for the 40-item test, the recommended pool should be at least a 300-item pool with the standard deviation (SD) of difficulty larger than the SD of the examinee's proficiency distribution. The required item-pool size for a CAT relies on the distribution of a student's proficiency population and the number of students (Reckase, 2010). However, it has typically been recommended that an item pool should be at least 10 to 12 times larger than the length of the CAT (Stocking, 1994). Along with the results of the overall adaptivity indices and the measurement precision of proficiency estimates, the performance of the proposed adaptivity measures provides more supportive evidence of Stocking's findings.

The amount of adaptation was closely associated with the standard errors in the proficiency estimation. Given the inspection of the relation between conditional adaptivity

measures and the standard errors of proficiency estimates, the DOD and ROI measures had strong and negative relations with the TSEM values. This suggests that good adaptation for CAT can lead to improving the measurement precision of proficiency estimates, and vice versa. This might be due to the fact that more precise proficiency estimates contribute to giving more appropriate information on the item-selection algorithm, leading to suitable customization of the items for a student's proficiency level. With TSEM and RMSE, however, a systematic pattern of CPRV was not found. In sum, the adaptivity of CAT should be closely associated with the measurement precision of proficiency estimates, but it was shown that they are different.

Second, the study demonstrated the practical utility of the proposed conditional adaptation statistics as diagnostic tools for improving the amount of adaptation for CAT in Research Question 2. The findings of the second simulation study indicated that the conditional adaptivity indices can provide insight into how to revise the existing item pool so as to improve the level of adaptation. Although the measurement accuracy and precision of the proficiency estimates were similar, the amount of adaptation could differ. Based on the initial computation of the adaptivity indices shown in Figure 5.17, a good example is the two proficiency regions of $-0.25 < \theta < 0.25$ and $1.75 < \theta < 2.25$.

As a result of adding to the existing item pool a series of fixed numbers of items, the information of which was high at each proficiency region, the more informative items available in the pool improved the level of adaptation at both regions. In particular, given the test length of 40 items and the composition of the item pool, the adaptivity indices were visibly improved and met the guidelines suggested in the first study when 30 informative items were added to the operational item pool at each region. To mimic a realistic situation, I did not fully control the quality of items in the master pool. That is, the master pool included items normally distributed,

implying that there might not be sufficiently high-quality items at each region. Hence, some of the added items were not as good as others for each region. The number of items to be added may depend on the quality of an item. For instance, an item with a high *a*-parameter to be administered at the proficiency level closer to its *b*-parameter can contribute a good deal to improving the adaptivity compared to low-quality items. With the high-quality items at a particular proficiency region, the smaller number of items may be needed to enhance the amount of adaptation for a CAT.

Third, the study in Research Question 3 examined how much "adaptation cost" occurred when the constraints of exposure control were imposed on the item-selection algorithm. The three exposure-control procedures considered in this study reported comparable measurement precision of proficiency estimates; nonetheless, the magnitude of adaptation was apparently different across the exposure-control procedures. The randomesque procedure did not compromise adaptation using either the regular or the optimal item pool. The Sympson-Hetter method, though, reduced the level of adaptation, especially for the DOD index. Interestingly, the *a*-stratified with *b*-blocking (BAS) approach contributed to improving adaptivity across the proficiency levels. Although this pattern would be consistent across different properties of item pools, the well-designed optimal item pool presented greater adaptivity across a broader range of proficiency levels for all exposure-control approaches of CAT. This suggests that, when exposure control is employed, good adaptation can occur with a good-quality item pool. It is additionally noted that the Sympson-Hetter method controlled well for overexposed items, whereas the BAS design showed more balanced-usage of items in the pool by controlling for underexposed items.

Fourth, another notable result of this research focused on the amount of adaptation presented by adaptive testing designs using the 3PL IRT model and the MLE for proficiency estimation. The specific designs considered here for a 40-item test were a fully item-level adaptive test (i.e., CAT) and a 1-2-3 three-stage multistage adaptive test (MST). It was shown that a MST reported obviously less adaptation than a CAT regardless of item-pool characteristics, though the two testing designs reported comparable accuracy of proficiency estimates in the moderate proficiency levels ranging from -1 to 2. The MST did not satisfy the guidelines of the three conditional adaptation indices across the proficiency levels. For the MST, the middle proficiency region showed relatively better adaptivity than other proficiency regions. An unanticipated result was that the MST formed from the optimal item pool did not present clearly higher adaptation than the MST created from the regular item pool. This differed from the CAT case. The CAT showed better adaptation over the full range of students' proficiency using the optimal item pool than it did when using the typical operational item pool.

Last but not least, empirical demonstration was made using real operational data from the NCLEX nursing licensure examination. Three conditional adaptivity measures indicated that this variable-length test was well designed, with a high-quality item pool for satisfying the purpose of the test showing good adaptivity, being near the cut-score of 0.0. That is, at the proficiency level closer to the cut-score, the three adaptivity indices met the benchmark values. This suggests that the test was very adaptive for classifying examinees into mastery and non-mastery of their proficiency for the nursing licensure—even when considering content balancing and exposure control. Overall, the proposed statistics were properly functioned as a diagnostic tool for understanding the amount of adaptation contingent on the proficiency continuum for an operational CAT.

6.2 Practical Utility of Conditional Adaptation Indices

The findings of the entire study reported that new conditional adaptivity measures were sensitive to item-pool characteristics, test designs, and test specifications that could affect the amount of adaptation and suggested some guidelines for interpreting the statistics. The study also provided evidence to support the usage of these conditional indices for helping test developers and measurement professionals revise test design or an item pool to improve the test adaptivity. This section presents a discussion of the practical utility of the conditional adaptation indices.

6.2.1 Diagnostic tools for improving adaptivity

The proposed adaptation indices can be used not only as quality control tools to monitor the adaptivity of an adaptive testing program, but also as diagnostic tools for improving adaptivity by revising an item pool or a test design. Practitioners may want to maintain the level of adaptation for adaptive tests across testing windows, item pools, subgroups of examinees, or time occasions. In practice, not all examinees can take the tests at the same time. Some may take the tests earlier than others, and some may take the tests through different windows or using a different item pool assembled from the master pool. Sometimes, test developers may be interested in inspecting a particular subgroup of examinees to determine if the tests adequately function well for the test purpose. Examinees may want to take a fair test that measures their proficiency as accurately and precisely as possible as well as a test with items well adapted for their proficiency levels. This desire may come to realization with adaptation statistics, as they can play a role in evaluating and tracking the amount of adaptation for the administered tests.

Beyond the quality control, the newly proposed adaptation indices are particularly useful in diagnosing a current test and to provide some directions for improving the adaptivity of the test by revising an item pool, test specifications, or test designs. A pivotal element of a CAT is

the item pool, such that the quality of the CAT is closely tied with the properties and the quality of the item pool (Flaugher, 2000). As demonstrated in Research Question 2, the level of adaptation can be improved by adding items that are informative at a proficiency region of interest to the existing item pool. Some test events can occur using multiple item pools even within the same session. To facilitate test fairness for students, the adaptation indices can be used, along with the item pool utilization index (Gönülates, 2015), to assess the performance of the tests with each item pool in such a way that the multiple item pools include sufficiently good items adapted for students as equally as possible.

The proposed adaptation indices can be used to determine the adaptive test design but also to revise test specifications that optimize the adaptivity over the proficiency continuum within a test design. As with the study connected with Research Question 4, measurement professionals and test practitioners can utilize the adaptation statistics to compare the performances of different test designs (e.g., linear fixed-test, item-level CAT, MST, and hybrid CAT [Wang et al., 2016]). Not only can this be done for individuals or subgroups of examinees using the conditional adaptivity indices but also for the entire sample of examinees using the overall indices.

Furthermore, the conditional adaptivity measures are particularly useful for modifying tests to improve adaptivity within an adaptive test design. A good example would be an MST. The amount of adaptation and measurement quality for the MST depends on how the MST is designed and structured using the available item pools. With the optimal item pool, for example, the MST did not outperform, as shown in Figure 5.27, the test created using the regular item pool. In this case, the composition of items for each module through stages can be modified, based on the resulting plot of adaptation indices, so as to enhance the adaptivity of the test for

high or low proficiency students. That is, if the goal of a test is to achieve equal measurement precision and adaptability for all students.

Regarding the case of an item-level CAT, the test specifications can be determined by comparing differences in the amount of adaptation. Such differences are based on constraints in the item-selection algorithm for content balancing, exposure control, and avoiding test speededness. Based on the resulting values of the adaptivity statistics, test practitioners can decide whether some constraints should be relaxed or whether items should be added to the item pool. Taken together, the new adaptation metrics suggested here contribute to gauging the potential improvement of adaptivity for CAT, conditional on an individual student's proficiency level or those of examinee subgroups. Of course, practical concerns in the testing situation and the purpose of the tests must be taken into account.

6.2.2 Use of conditional adaptation indices in automated test assembly

The proposed conditional adaptation indices can be used as a constraint or as an objective function to assemble adaptive tests from a given item pool. Based on van der Linden's (1998b) distinction of test specifications in automated (optimal) test assembly, constraints are a test attribute or a function of item attributes that need to be met by setting an upper and/or a lower limit. Meanwhile, objective functions are the attribute(s) to be optimized by attaining a minimum or maximum value. The test assembly algorithm can be defined in numerous ways using the proposed conditional indices. With a constraint, for instance, a lower limit of each of the three adaptation measures can be set so that all students receive the test items yielding at least a certain value of the lower limit for each adaptation index. In a similar vein, the adaptivity indices can serve as an objective function of automated test assembly. They do so by setting the target values of the three adaptation indices so as to assemble, while satisfying all of the test constraints, the

test items customized for students' proficiency levels. The target values may be identified as the maximum mean value of each adaptation measure through preliminary analyses using sets of field tests or a small simulation study.

6.3 Alternative Ways to Define Conditional Adaptation Indices

The amount of adaptation can be quantified differently depending on how its concept is operationally defined. This study quantifies the amount of adaptation based on concepts of a highly adaptive test and provides well-customized items for the examinees' proficiency levels. To measure the amount of adaptation, this study establishes three quantities: (1) the DOD index, (2) the CPRV index, and (3) the ROI index. The overall adaptation indices (Reckase et al., 2018) focused on whether the characteristics of the administered items were well matched for students' final proficiency estimates. The conditional adaptivity quantities proposed here, in contrast, focus more on how well a CAT uses—during its administration—an available item pool and item-selection algorithm to give the items adapted for an individual's provisional proficiency estimate. The latter is beneficial in that the final proficiency estimates cannot be known during the CAT process. Furthermore, the current proficiency estimates can assess whether the item-selection algorithm is working correctly so as to provide a well-matched item during the CAT process even though the interim estimate may deviate from the final estimates. It was found, via comprehensive simulation and empirical studies, that the proposed three indices functioned as expected. Nonetheless, it is still necessary to discuss here alternative ways to define the current conditional adaptation indices.

Rather than the usage of interim proficiency estimates, the conditional adaptation indices can be computed using the examinee's final proficiency estimates. In this case, the adaptation indices are conceptually more associated with seeing whether the optimal set of items are

administered to each examinee, assuming that their final proficiency estimates are not biased and sufficiently close to their true proficiency levels. Taken as an example is the DOD index. With the *final* proficiency estimate, the DOD can evaluate whether the items were properly administered for matching their location (either item difficulty or the location of maximum information) to the examinee's final proficiency estimate. So, the deviation of the item location from the final proficiency estimate can be large early on in the test, more likely yielding a DOD value below 0.

Another example is the ROI index. The current ROI index compares the information function of the administered item at the provisional proficiency estimate to the potential maximum information that the item can reach. Instead, the observed item information, or the numerator of the ROI quantity, can be computed at the final proficiency estimate. This might allow for knowing the amount of information that the test provides to individual examinees around their final proficiency estimate. It can be blind, though, to whether the items were, during the intermediate stages of the CAT process, appropriately presented and well utilized for the examinees.

Moreover, the ROI index can have a differently defined criterion value, which is placed in the denominator of the index. This study identifies the optimal criterion value of the information as the maximum potential information that an item can have. However, the optimal information can be determined through one of two ways. It can be through the maximum information available in an existing item pool—the theoretical limit of the maximum information given the Rasch model. Or it can be through the maximum information that can be obtained from the most informative items in the pool at the true or the final estimated proficiency levels. In such cases, the ROI would not evaluate how much of an item's potential information is realized

at the interim proficiency estimate during the CAT. It would instead assess whether a perfect item (that matches well the true or final estimated proficiency) is presented to each examinee using the current item pool (e.g., Gönülates, 2015). For instance, the percent of information index (Kingsbury & Wise, 2018) compared the observed actual test information to the optimal information obtained by administering a test at the true proficiency level. Although the ROI seems to have different conceptualizations depending on how the optimal information is identified, these conceptualizations still have something in common—they are information-based measures.

6.4 Implications

Overall, with the help of guidelines and their own understanding of the concepts, researchers and practitioners could easily interpret the three conditional adaptivity indices investigated in this study. These new measures allow us to understand how much adaptation of a given test occurs across proficiency levels or subgroups of students. They also help us understand the "adaptation costs" resulting from item pool characteristics, adaptive test designs, constraints on the item selection, among other test specifications. Together with the overall adaptivity measures for the entire groups of examinees, the newly proposed conditional measures for the amount of adaptation offer unique contributions. Indeed, they enable practitioners to have a better sense of which adaptive test may not be very adaptive for individual examinees or particular subgroups of examinees.

Another benefit of these indices are the various ways in which one can summarize the distribution of conditional adaptivity measures. The most intuitive way is to visualize, as I did in chapter 5, the values of the adaptation statistics against the students' entire proficiency continuum or against particular proficiency regions of interest using a scatter plot, histogram, or

box plot. As with the overall adaptation statistics, the distribution of the conditional adaptation measures can also be summarized using the descriptive statistics such as a mean, median, and standard deviation of the values for a set of tests administered to a group of examinees in the population. If the distribution is skewed, the mean and median values of the statistics are discrepant. It would be advisable here to look at the entire distribution of the conditional measures using a graphical method. One thing that has to be noted here is that the amount of adaptation using the distribution of the adaptivity indices should be understood given the specific goals of a testing program. Although some examinees received items that were poorly customized for their proficiency levels, this might, depending on the testing purpose, matter little.

The concept for the amount of adaptation should be understood differentially from that of measurement accuracy and precision, despite their being tightly associated with each other. The relations of the standard errors in proficiency estimates with each of three adaptivity indices, were not, as we noted above, perfectly correlated to one another; even the CRPV index did not show the apparent pattern with the measurement quality. Although students had similarly acceptable levels of measurement quality of their proficiency estimates, the quality of adaptation could differ, as seen in Research Question 2, depending on numerous issues. These include the size and characteristics of an item pool, the functioning of the item-selection algorithm, test design, and other issues that can affect adaptivity of CAT. That is, the proficiency was accurately and precisely estimated to some extent, but the test could not provide best items adapted to the examinees because of deficiencies in the item pool or some problems in the item-selection algorithm.

The tests, in contrast, were very adaptive. The quality of measurement, however, could be poor. This statement is evidenced, for instance, by the findings of the comparison of the conditional adaptivity values between the MLE and the EAP in Research Question 1. In general, the EAP estimator showed the greater adaptivity for the 40-item test than the MLE over the proficiency continuum. The proficiency estimates obtained using EAP, however, were biased. This bias is due to the property of the EAP regressing the proficiency estimates toward the mean of a prior distribution (Ho & Dodd, 2012; Kim & Nicewander, 1993). Students took the items that were well adapted for their current proficiency estimates, but this will not accordingly reduce the bias caused by the property of proficiency estimator. Additionally, it may be imagined that with a very short test, the level of adaptation could be high with the well-designed item pool and test specifications, but the proficiency would be poorly estimated because the number of items might not be adequate for a precise and accurate estimation of the proficiency. Therefore, the high adaptivity of a CAT does not always guarantee the high measurement efficiency of proficiency estimates.

6.5 Limitation and Future Research

This section briefly discusses some limitations of the study while also offering directions for future research. First, reported findings from this study examined the differences in the amount of adaptation affected by limited aspects. While the concepts and measures to evaluate the level of adaptation have recently received growing attention, more research is needed in this area. Future study should elaborate on the current adaptation measures by examining their performance under other factors that plausibly affect the adaptivity. These factors include not only other adaptive testing designs (e.g., hybrid CAT design) and test specifications such as latency-constraints for preventing test speededness and content-constraints for content balancing,

but also using other real operational adaptive testing data with a different purpose of a test (e.g., equal measurement precision over the proficiency continuum). Researchers may come up with further alternatives to the existing adaptivity indices with different assumptions and definitions of "the amount of adaptation."

Second, in Research Question 2, the quality of the items to be added from the master pool were not fully controlled. The current study found that to visibly improve the level of adaptation from all three aspects called for approximately 30 items. However, the number of items needed to improve the level of adaptation can differ, depending on the quality of an item and the location in which that item is efficiently presented. What are the features of a high-quality item? According to Eignor, Stocking, Way, & Steffen (1993), a high-quality item is one that is the most informative at a student's current proficiency estimate. Given the Fisher information function, the amount of the information is maximized when an item has a high *a*-parameter and *b*-parameter close to the examinee's true proficiency level (Hambleton & Swaminathan, 1985). The approach in this study requires multiple iterative procedures to improve the amount of adaptation given the distribution of test takers. These procedures are selected by checking the statistics after the removal or addition of proper items from the available item pool. Occasionally, this iterative approach may be time-consuming in real operational settings. Future research can explore possible ways of coming up with an equation that computes how many items need to be added, taking into account the amount of information that an item has at the current proficiency estimate. With the help of that equation, the item pool or test designs are expected to be more efficiently revised by adding the items necessary to improve adaptivity, in lieu of the iterative procedure taken in this study.

Third, the findings gleaned from Research Question 4 underscores the paramount importance of optimally designing MST modules. To obtain more precise proficiency estimates, the literature has mostly focused on either MST path structures or the stage and module configurations (e.g., Luo & Kim, 2018; Patsula, 1999; Xiong, 2018; Zenisky, 2004). Relatively little is known, though, about how the ways of assembling the modules and composition of item characteristics affect the proficiency estimates and the level of customization for MST given the examinees' proficiency distribution. Future research could address this concern by exploring other ways of optimally designing MST beyond the approach of using the target TIF.

Fourth, little research has yet explored the measures of the amount of adaptation in the mixed-format adaptive testing context. Mixed-format tests that include multiple-choice (MC) items, constructed-response (CR) items, and testlets have been commonly used in educational large-scale testing (Kim, Walker, & McHale, 2010; Kuechler & Simkin, 2010; Yao & Schwarz, 2006). Due to their enhanced psychometric features, mixed-format tests can be more informative, efficient, and valid as well as more promising for future applications implementing innovative items (e.g., technology-enhanced items, multiple-response items, hot-spot items) in CAT programs. This results in enhanced content coverage and measurement accuracy (Jiao, Liu, Haynie, Woo, & Gorham, 2012; Wendt, 2008). It is recommended to polytomously score and calibrate such innovative items, CR items, and testlets using the polytomous IRT models in CAT programs (e.g., Jiao et al., 2012). It would be interesting for future research to investigate whether the existing adaptation measures could work well with the polytomously scored items when their item parameters are calibrated using the polytomous IRT models with the overall item difficulty parameterization. Also, existing studies have explored the influences of test designs using only dichotomous items and without consideration of content balancing. Multiple item

types and content balancing may produce more variations in the test designs, resulting in different levels of adaptation. This calls into question how much adaptation occurs as a result of the mixed-format test designs with content balancing approaches.

Lastly, this dissertation suggests some guidelines to interpret the indices based on the current simulations results, but the benchmarks were not solely evaluated yet. Hence, further research is needed to work on elaborating the benchmark values for the adaptation indices via a Type I error study and a power study. Additionally, the dissertation evaluated the performance of the adaptation indices using the maximum Fisher information (MFI) item selection procedure. In fact, the proposed measures were developed based on the MFI item selection criterion. In literature and operational CAT programs, there are numerous item selection criteria out there (see Section 2.2.2). It is worthy of exploring whether the adaptation measures properly work with other item selection procedures such as b-matching and Bayesian criteria in future research.

A promising testing format has emerged in recent years for the next-generation assessments; that format is adaptive testing. In fact, numerous testing programs have already employed a variety of adaptive tests. It is now time to reconsider and evaluate how adaptive the current tests are and how the tests can be improved so that they rise to the purpose of "adaptive" tests. It is also strongly recommended that adaptive testing be implemented knowing the following: such testing is attended with psychometric impacts from the test designs and specifications on adaptivity. This is particularly true in a situation where a testing agency has just started transitioning its testing format from the paper-and-pencil linear testing to the adaptive testing. In the area of adaptive testing, therefore, the newly established adaptation indices in this study can direct test practitioners toward consequential consideration of improving the item pools and test designs for adaptive tests.

APPENDIX

APPENDIX

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 1

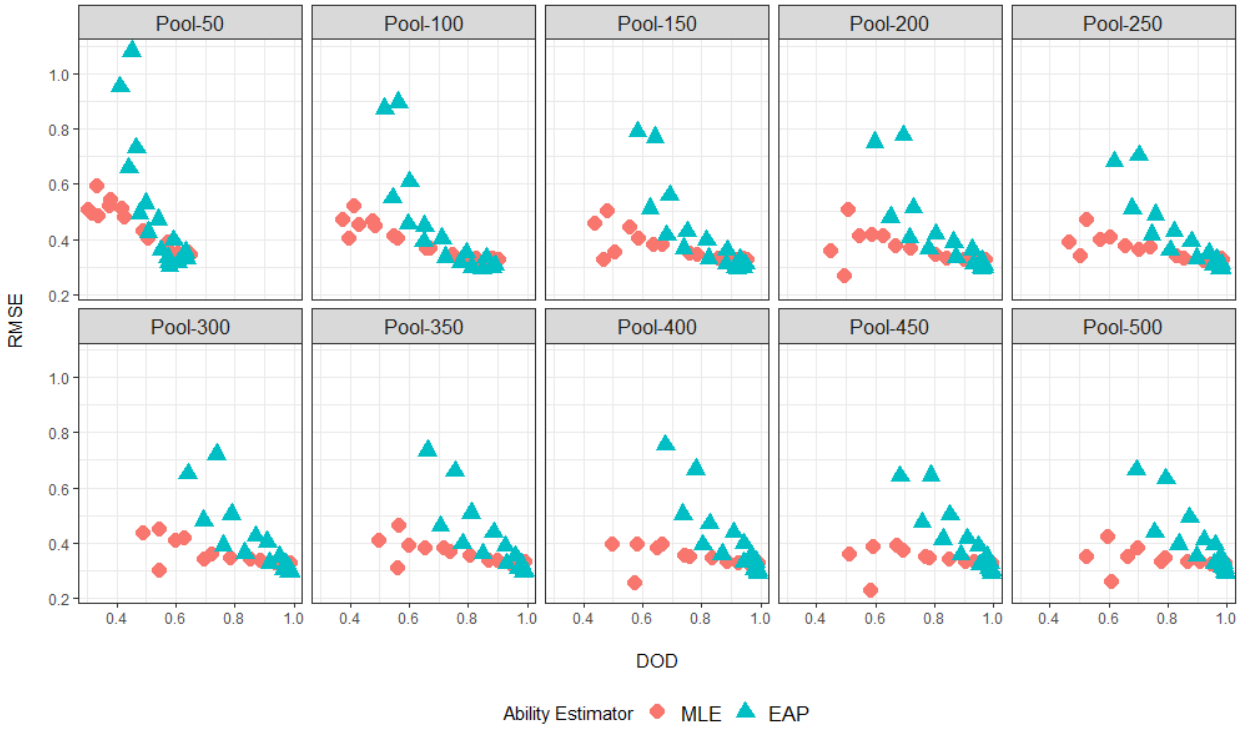
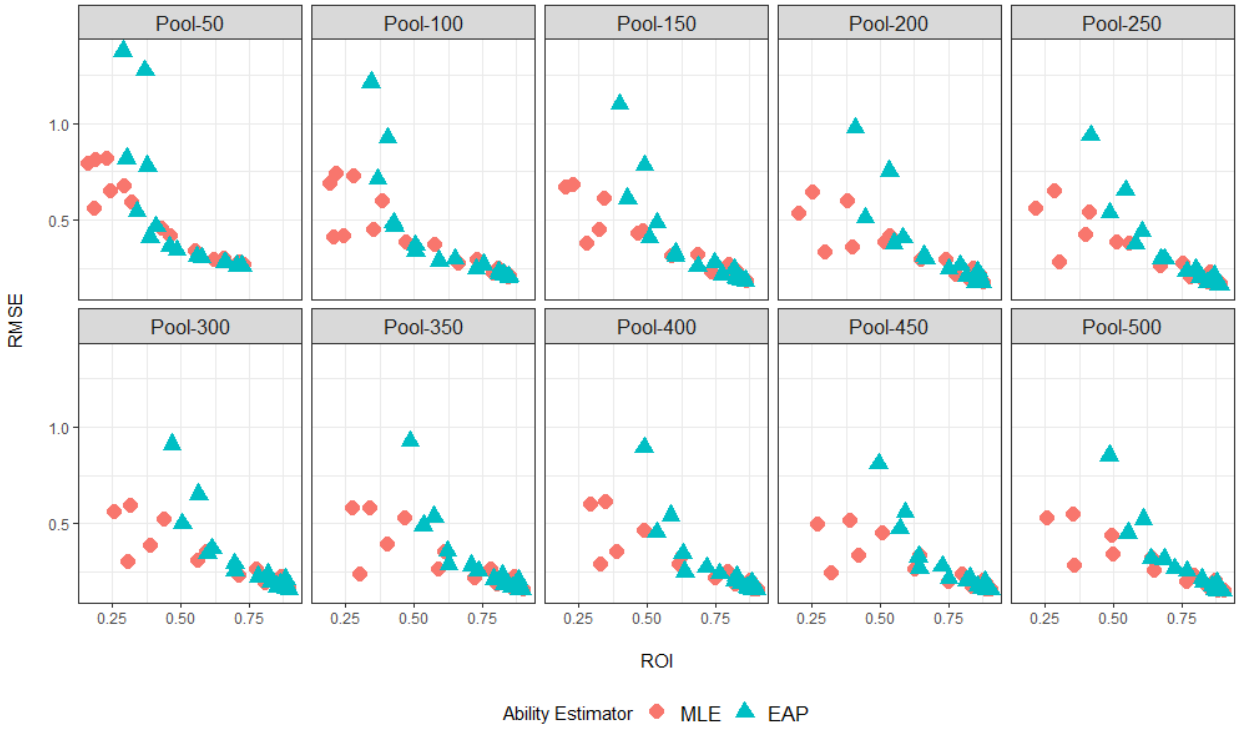


Figure A.1. Relationship of RMSE with conditional adaptivity indices (DOD, CPRV, and ROI) for a Rasch-based CAT by item pool size and proficiency estimator

Figure A.1. (cont'd)

Figure A.2. Relationship of RMSE with conditional adaptivity indices (DOD, CPRV, and ROI) for a 3PL-based CAT by item pool size and proficiency estimator

Figure A.2. (cont'd)



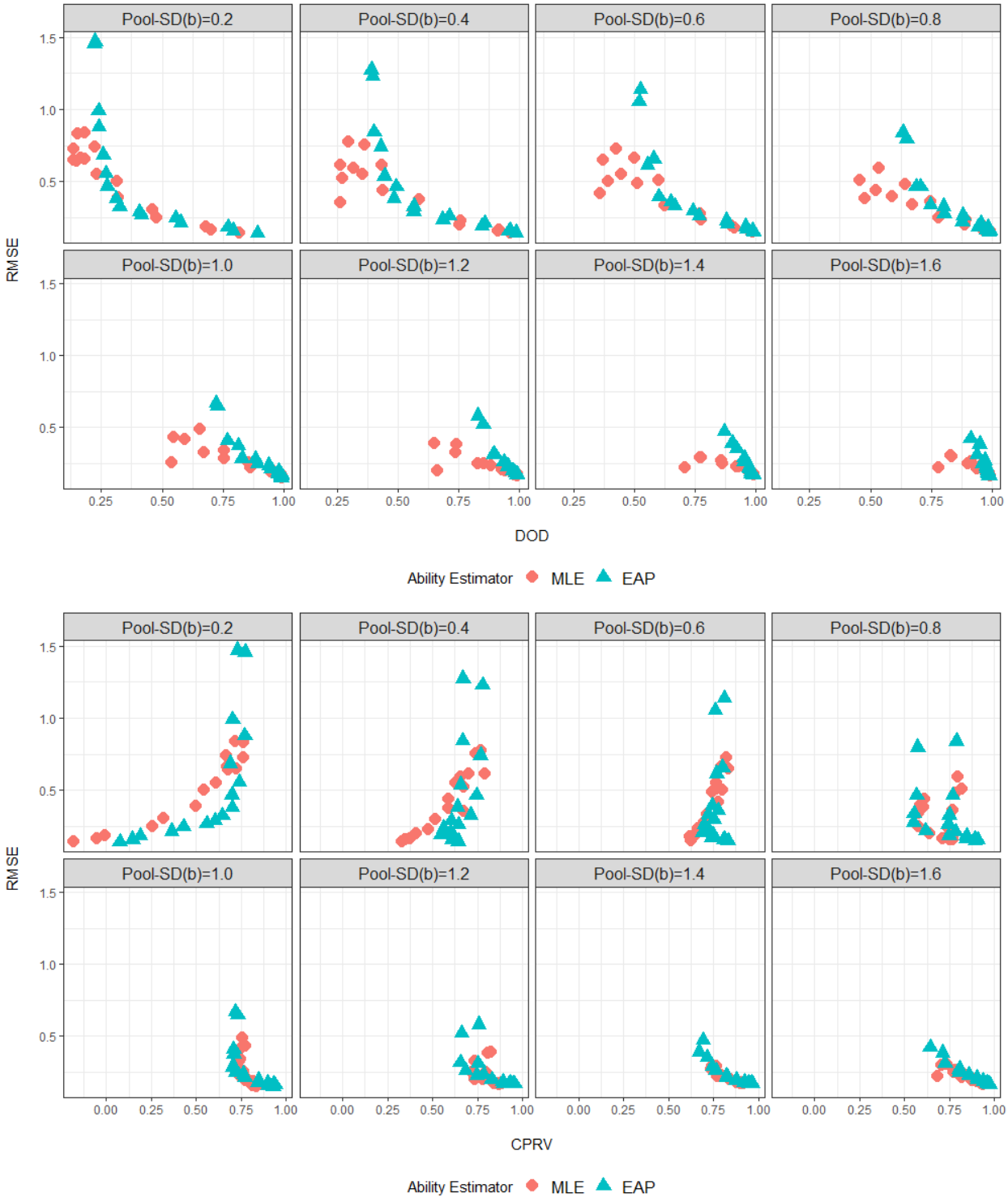
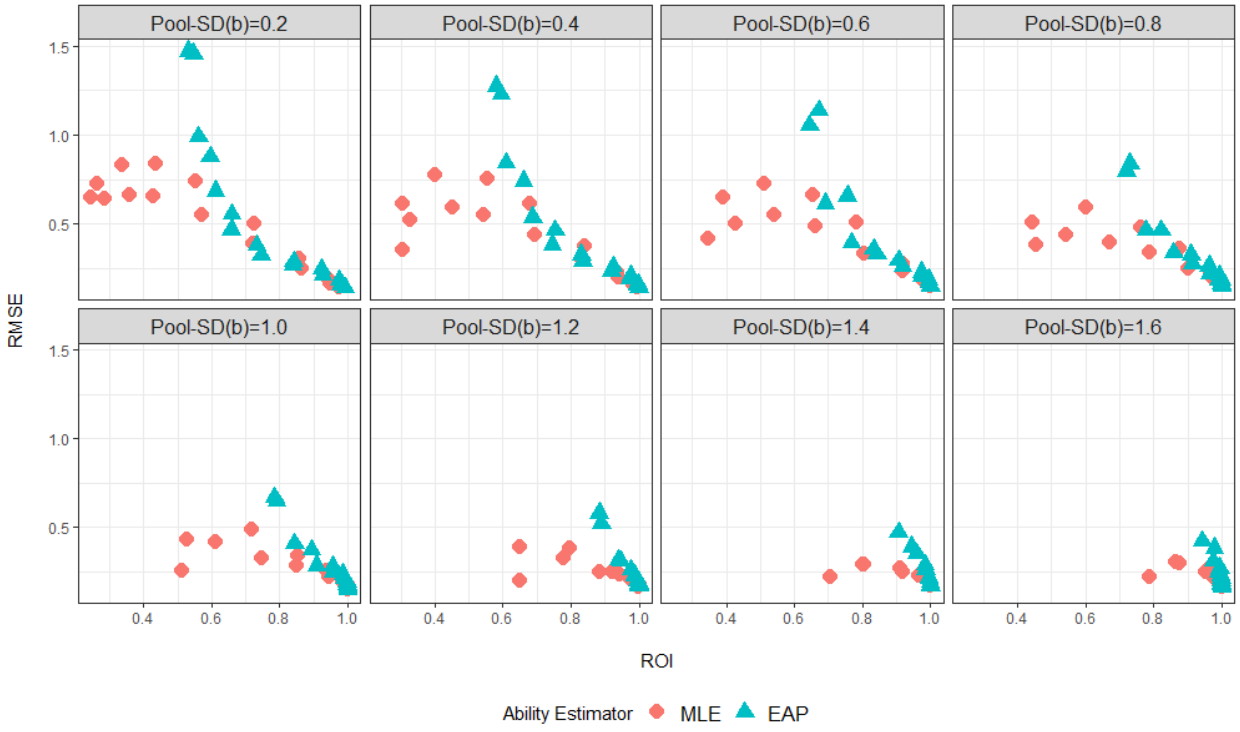


Figure A.3. Relationship of RMSE with conditional adaptivity indices (DOD, CPRV, and ROI) for a Rasch-based CAT by item pool spread and proficiency estimator

Figure A.3. (cont'd)



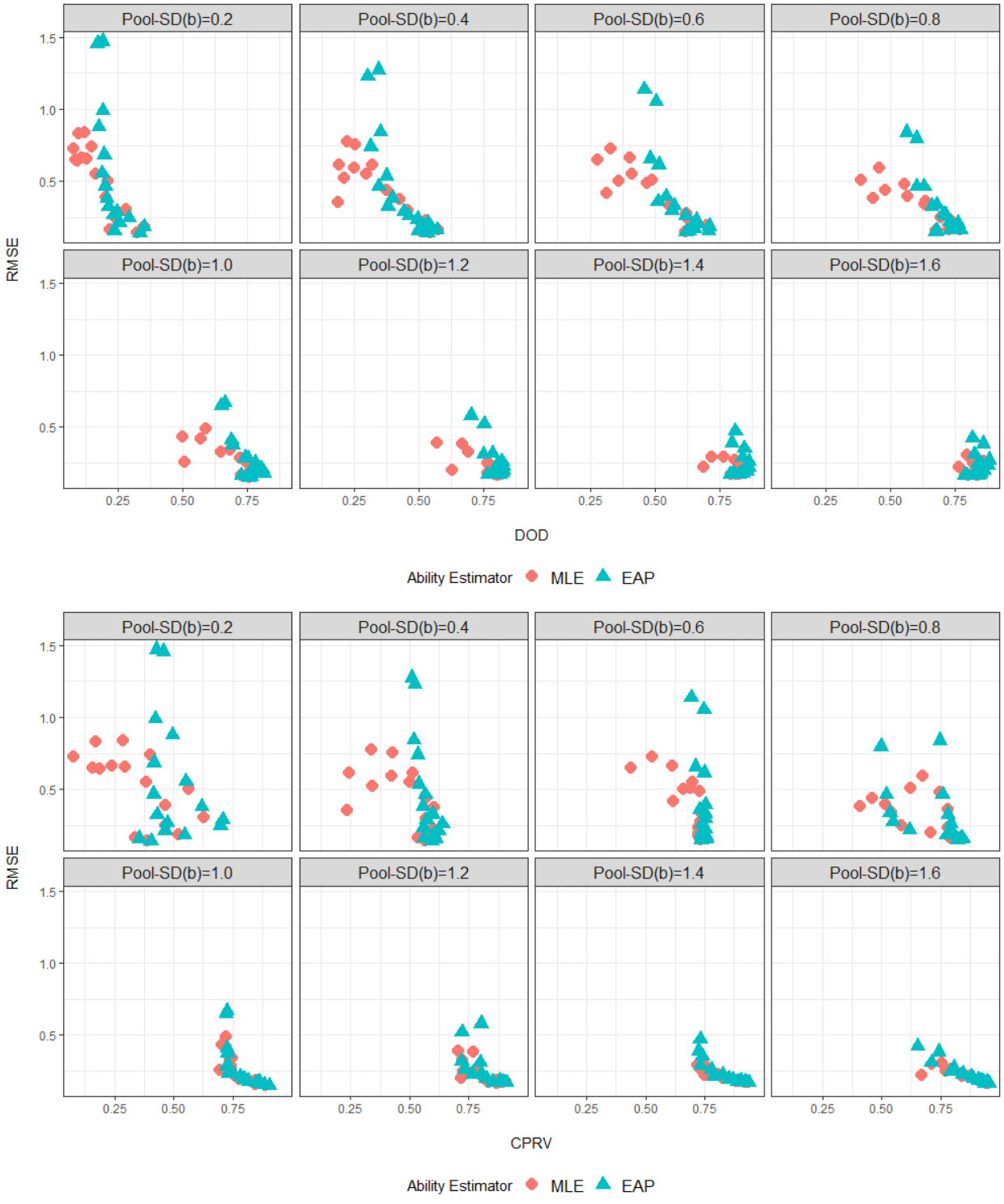
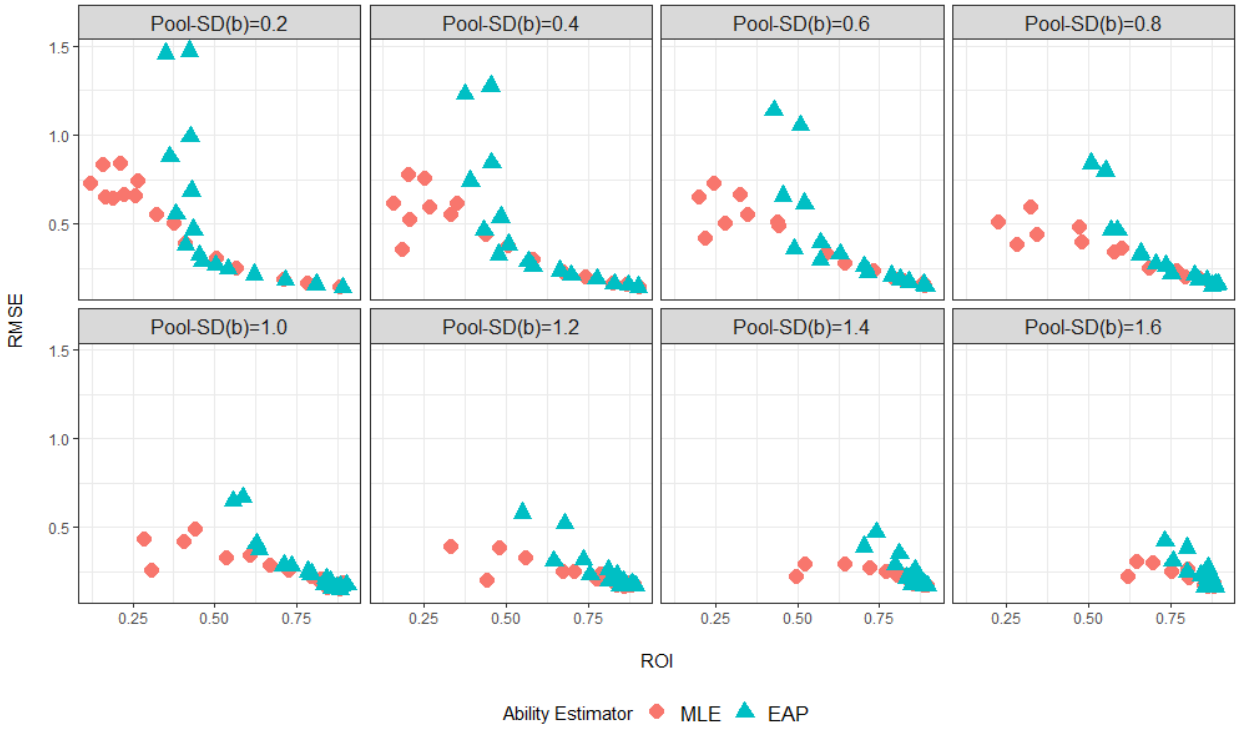


Figure A.4. Relationship of RMSE with conditional adaptivity indices (DOD, CPRV, and ROI) for a 3PL-based CAT by item pool spread and proficiency estimator

Figure A.4. (cont'd)



REFERENCES

REFERENCES

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137-149.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 374-472). Reading, MA: Addison-Wesley.

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

Chang, H.-H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methods for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage.

Chang, H. H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25*, 333-341.

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262–274.

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

Chang, H. H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H (2000). A compassion of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

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.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*, 419-440.

Chang, H.-H., Qian, J. & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25*, 333-341.

Conley, T. D. (2018). The Promise and Practice of Next Generation Assessment. Cambridge, MA: Harvard Education Press.

Davey, T. (2005, April). *An Introduction to bin-structured Adaptive Testing*. Presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (ETS Research Report No. 93-56). Princeton, NJ: Educational Testing Service.

Embretson, S.E. (2001). The second century of ability testing: Some predictions and speculations. Retrievable at http:// www.ets.org/Media/Research/pdf/PICANG7.pdf.

Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, … D. Thissen (Eds.), *Computerized adaptive testing A primer* (2nd ed., pp. 37-60). Mahwah, NJ: Lawrence Erlbaum.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5, 4-38.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Gönülates, E. (2015). *A novel approach to evaluate item pools: The item pool utilization index* (Doctoral Dissertation).

Gu, L. & Reckase, M. D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston, MA: Kluwer-Nijho Pub.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE.

Han, K. T. (2012). An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement, 49*, 225-246.

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement, 40*, 289–301.

He, W., & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement, 74,* 473-494.

Hembry, I. F. (2014). *Operational characteristics of mixed-format multistage tests using the 3PL testlet response theory model* (Doctoral Dissertation).

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153-160.

Ho, T.-H., & Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed-format adaptive test. *Applied Measurement in Education, 25,* 305–326.

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement, 72,* 493-509.

Ju, U. & Lee, Y. (2018, July). *Effects of ability estimation methods on the amount of adaptation for computerized adaptive tests.* Paper presented at the biannual meeting of the International Test Commission, Montreal, Canada.

Kim, J. K. & Nicemander, W. A. (1993). Ability estimation for conventional tests. *Psychomtrika, 58,* 587–599.

Kim, S., Ju, U., & Reckase, M. D. (2018, April). *Evaluating indicators of amount of adaption to 3pl computerized adaptive test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47,* 36-53.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

Kingsbury, G. G. & Wise, S. L. (2018, July). *A new measure of adaptation based on test information*. Paper presented at the biannual meeting of the International Test Commission, Montreal, Canada.

Kuechler, W.L., & Simkin, M.G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8,* 55-73.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model* (Doctoral dissertation).

Luo, X. (2015). *Incorporating mixed item formats in CAT: A comparison of shadow test and bin-structured approaches* (Doctoral Dissertation).

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement, 55*, 243–263.

Luecht, R. M. & Clauser, B. E. (2002). Test models for complex CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum.

Luecht, R.M. & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Mao, L. (2014). *Designing p-Optimal Item Pools for Multidimensional Computerized Adaptive Testing* (Doctoral Dissertation).

McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*, 121-140.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224-236). New York, NY: Academic Press.

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006, April). *A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based on the 3PL and partial credit models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Meijer, R. & Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304.

Minnesota Department of Education (2017). *Technical manual for Minnesota standards-based accountability and English language proficiency assessments: For the academic year 2015–2016.* Minnesota Department of Education. Roseville, MN.

National Council of State Boards of Nursing (2016). *NCLEX-RN examination: Test plan for the National Council Licensure Examination for Registered Nurses*. National Council of

State Boards of Nursing. Chicago, IL. Retrieved from https://www.ncsbn.org/RN_Test_Plan_2016_Final.pdf

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* (Doctoral Dissertation).

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 119-141). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer Verlag.

Patience, W. M., & Reckase, M. D. (1980, April). *Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Doctoral dissertation).

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1-20.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321-333). University of California Press Berkeley, CA.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria. URL https://www.R-project.org/.: R Foundation for Statistical Computing.

Reckase, M. D. (1975, April). *The effect of item choice on ability estimation when using a simple logistic tailored testing model*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Reckase, M. D. (1976, April). *The effect of item pool characteristics on the operation of a tailored testing procedure.* Paper presented at the spring meeting of the Psychometric Society, Murray Hill, NJ.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice, 8*, 11-15.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*, 127-141.

Reckase, M. D., Ju, U. & Kim, S. (2017, November). *Differences in the amount of adaptation exhibited by various computerized adaptive testing designs*. Paper presented at the 17th Annual Maryland Assessment Research Center (MARC) conference, Maryland.

Reckase, M. D., Ju, U. & Kim, S. (2018). Some measures of the amount of adaptation for computerized adaptive tests. In M. Wiberg, S. Culpepper, R. Janssen, J. González, D. Molenaar (eds.), *Quantitative Psychology, IMPS 2017, Springer Proceedings in Mathematics & Statistics 233* (pp. 25-40). Charm: Springer.

Reckase, M. D., Ju, U. & Kim, S. (2019). How adaptive is an adaptive test: Are all adaptive tests adaptive? *Journal of Computerized Adaptive Testing*, 7, 1-14.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society.

Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Pearson Research Report. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/99A4327B-5968-4AB2-A8CD-8D502D22C2DE/0/WeightedPenaltyModel.pdf

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education, 19*, 257-260

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (ETS Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the Annual Meeting of the Military Testing Association. Navy Personnel Research and Development Center: San Diego, CA.

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

The MathWorks, Inc. (1984-2015). MATLAB version 10.1. Natick, Massachusetts: The MathWorks Inc.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, … D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Vale, C. D. & Weiss, D. J. (1977). *A rapid item-search procedure for Bayesian adaptive testing* (Research Report 77-4). Minneapolis, MN: University of Minnesota, Psychometric Methods Program, Adaptive Testing Laboratory.

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W. J. (1998b). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195–211.

van der Linden, W. J. (2010). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing* (pp. 31-55). New York, NY: Springer.

van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31*, 81-99.

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. (2000). Designing item pools for computerized adaptive testing. In van der Linden, W. J. & Glas, C. A. W. (Eds.) (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer.

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor item response theory Approach. *Frontiers in Psychology, 7*.

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement, 53*, 45-62.

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

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

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Way, W., Zara, A., & Leahy, J. (1996, April). *Modifying the NCLEX^{TM} CAT item selection algorithm to improve item exposure*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*, 1-27.

Wendt, A. (2008). Investigation of the item characteristics of innovative item formats. *Clear Exam Review*, *19*, 22-28.

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*, 21-30.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135-155.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*, 5-21.

Xiong, X. (2018). A hybrid strategy to construct multistage adaptive tests. *Applied Psychological Measurement, 42,* 630-643.

Yan, D. von Davier, A. A. & Lewis, C. (Eds.) (2016). *Computerized Multistage Testing: Theory and Applications.* Boca Raton, FL: CRC Press.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 37*, 3-23.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Doctoral Dissertation).

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage adaptive testing for a large-scale classification test: the designs, automated heuristic assembly, and comparison with other testing modes* (ACT Research Report 2012-6). Iowa City, IA: ACT Inc.

Zhou, X., & Reckase, M. D. (2014). Optimal item pool design for computerized adaptive tests with polytomous items using GPCM. *Psychological Test and Assessment Modeling*, *56*, 255-274.