A NOVEL APPROACH TO EVALUATE ITEM POOLS: THE ITEM POOL UTILIZATION INDEX

By

Emre Gönülateş

A DISSERTATION

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

Measurement and Quantitative Methods – Doctor of Philosophy

2015

ABSTRACT

A NOVEL APPROACH TO EVALUATE ITEM POOLS: THE ITEM POOL UTILIZATION INDEX

By

Emre Gönülateş

In this study, an index to quantify the adequacy of an item pool of an adaptive test for a given set of test specifications and examinee population is introduced. This index is called the Item Pool Utilization Index (IPUI). The IPUI ranges from 0 to 1, with values close to 1 indicating the item pool can provide optimum items to examinees throughout the test. This index can be used to compare different item pools or diagnose the deficiencies of a given item pool by quantifying the amount of deviation from a perfect item pool.

Simulation studies were conducted to evaluate the capacity of this index for detecting the inadequacies of both simulated and operational item pools. The added value of this index was compared to the existing methods of evaluating the quality of computerized adaptive tests (CAT).

Results of the study showed that the IPUI can detect even slight deviations of the item pools from an optimal item pool. It can uncover the shortcomings of an item pool that other outcomes of CAT cannot detect. Additionally, it can be used to diagnose the weaknesses of the item pool and guide test developers to improve their item pools.

Keywords: Computerized Adaptive Test, Item Pool, Item Pool Design

I dedicate this dissertation to my wife Funda, my children Meryem, Bilge and Elif, and my parents, Meryem and Selahattin.

ACKNOWLEDGEMENTS

This dissertation thesis is the outcome of many years of studies at Michigan State University. The idea for this dissertation came from an advanced psychometry course I took from Dr. Mark Reckase back in Fall semester of 2013. At that time, I was helping Dr. Reckase in developing ideal item pools for a testing company. When I saw the item selection algorithm introduced in Han (2012), which was one of the texts we read in the course, the idea of this thesis was born.

First, my deepest thanks goes to my academic advisor and committee chair Dr. Mark Reckase. During my academic life at MSU he has been a wise, thoughtful and kind mentor. He was generous in sharing his wisdom and knowledge. He has been a great example of a hardworking, dedicated and passionate scholar. He introduced me to many projects that taught me both practical and theoretical aspects of our field. He spent many hours with me to cultivate the idea of this thesis into a finished work that I'm very proud of.

National Council of State Boards of Nursing (NCSBN) provided financial support at the early stages of this dissertation. The organization also provided operational data used in this study that helped me to show the practical use of the index. I want to thank Drs. Ada Woo, Hong Qian and Doyoung Kim from NCSBN for their help and support.

I wish to thank my friends and colleagues, Eun Hye Ham, Nazlı Uygun, Liyang Mao, Xin Luo, Chi Chang, Francis Smart, Ifeoma Iyioke, Bing Tong, Anne Traynor, Lihong Yang, Tingqiao Chen, Keyin Wang, Hyesuk Jang and Xuechun Zhou who enrich my life at MSU and contributed to my learning.

I'm grateful to my dissertation committee members, Drs. Spyros Konstantopoulos, Richard Houang and Christopher Nye for their willingness to review my work. Their feedback improved this dissertation immensely. In addition, I'm thankful to the faculty at Measurement and Quantitative Methods Program, especially Ken Frank, Tenko Raykov, Kimberly Maier,

Bill Schmidt. They helped me to establish the basis for my advanced studies. I'm grateful to the support of Dr. Ed Roeber. He advised me in the first two years of my studies at MSU and helped me to find my way in this large field of study.

My unique gratitudes goes to my parents, Meryem and Selahattin Gönülateş, for their sacrifices on my behalf and their unconditional love. Their endless support and encouragement throughout my life has helped me accomplish this and many other goals in life. I'm grateful to my sister Zeynep and brother Ahmet Talha for being there whenever I need them.

My children, Meryem, Bilge and Elif, they are the joy of my life. Over many hours they put up with an often absent and distracted father. I thank them for their love and the happiness they bring to my life.

And finally, I owe this dissertation to the unwavering support of my wife, Funda. Without her, I would not have been able to complete this degree. During these years she has a lot of things on her shoulders. She gave birth to three wonderful children and raised them, pursued a PhD of her own, involved in many projects and much more. But still, she always make sure I have enough space and time for my work. Her moral support helped me to find my way during the gloomy days and she shared my joy in many happy days. She spent many hours reading my drafts and gave valuable feedback. I'm thankful to her for encouraging me to follow my dream and believe in myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

1PL one-parameter logistic. 4, 6, 27, 33–35, 37, 40, 162–164

2PL two-parameter logistic. 6, 27, 34, 40, 41, 162–164

3PL three-parameter logistic. 6, 9, 27, 34, 40, 41, 163, 164

CAT computerized adaptive testing. 1–3, 5–10, 12–16, 18–24, 26, 28–32, 36, 37, 39–53, 60–62, 72, 73, 79, 84, 86, 93, 96, 102, 111, 131, 138, 141–151, 154–157, 160, 161, 163–165

EAP expected a posteriori. 16, 18, 37, 154

IPUI item pool utilization index. 29–35, 38–46, 49–54, 57–62, 65, 66, 68–70, 73, 74, 80–84, 86, 94–96, 99–103, 105, 106, 111–115, 117–120, 124, 130, 131, 136–144, 146–165, 194

IRT item response theory. 5, 16, 27, 28, 163–165

MAP maximum a posteriori. 18, 154

MCAT multidimensional computerized adaptive test. 164

MFI maximum Fisher information. 8–10, 37, 155, 162

MIRT multidimensional item response theory. 13, 164

MLE maximum likelihood estimation. 7, 16–20, 37, 48

MSE mean squared error. 23, 30, 36, 42, 44, 46, 52, 54, 56, 57, 60, 62–64, 73, 74, 78, 86, 89, 90, 94, 112, 124, 134, 138, 149

NCLEX-RN National Council Licensure Examination for Registered Nurses. 47–50, 52, 107, 124, 127, 130, 131, 144, 153, 159

NCSBN National Council of State Boards of Nursing. 47, 49, 107, 143

P&P paper and pencil. 6, 12, 21, 22

RMSE root mean squared error. 4

SE standard error. 9, 10, 23, 30–33, 36, 42, 44, 46, 52, 54–58, 60–63, 68, 69, 73, 74, 77, 78, 84, 86, 88, 89, 94, 96, 98–100, 112, 120–124, 133–135, 138, 142, 146, 149, 151, 154, 156–158

SEM standard error of measurement. 4

TIF test information function. 25

CHAPTER 1

INTRODUCTION

Increasing availability of computers and relative advantages of computerized adaptive testing (CAT) over paper based tests boosted the usage of CAT in recent decades. Mainly, a CAT enables more efficient measurement of examinee abilities, shorter test lengths, and more precise ability estimates for examinees at the extreme ends of the ability distribution. But these benefits come with costs. Among others, the requirement for a large item pool is the most challenging one. An item pool is the collection of items that will be used to construct individual adaptive tests for examinees. An item pool should include sufficient number of high quality items that are targeted to the examinee population (Parshall, Spray, Kalohn, & Davey, 2002). It should meet the content specifications of the test and provide sufficient information at all levels of the ability distribution of the target population (van der Linden, Ariel, & Veldkamp, 2006). Flaugher (2000) underlined the importance of item pools in a CAT:

> Obviously, the better the quality of the item pool, the better the job the adaptive algorithm can do. The best and most sophisticated adaptive program cannot function if it is held in check by a limited pool of items, or items of poor quality. (p. 38)

To demonstrate the importance of a large item pool, a very basic adaptive test has been simulated. Figure 1.1 shows the CAT progress of two examinees for a test with same test specifications and item pool. The item pool consists of 50 items with item difficulties generated from a standard normal distribution. The points in the figure show the intermediate ability estimates of the examinees. Blue 'b' points represent the item difficulty parameters that are administered to the examinees at each stage of the adaptive test. True ability parameters of

Examinee 1 and Examinee 2 are 0 and 1.5, respectively. For the first examinee, the item pool is very suitable. At each step, the item pool can provide an item with a difficulty parameter which is very close to the Examinee 1's intermediate ability estimate. On the other hand, for Examinee 2, after correctly answering a couple of questions, the item pool is out of difficult items. As a result, even though Examinee 2 correctly answers each item, the CAT algorithm presents easier items to this examinee.

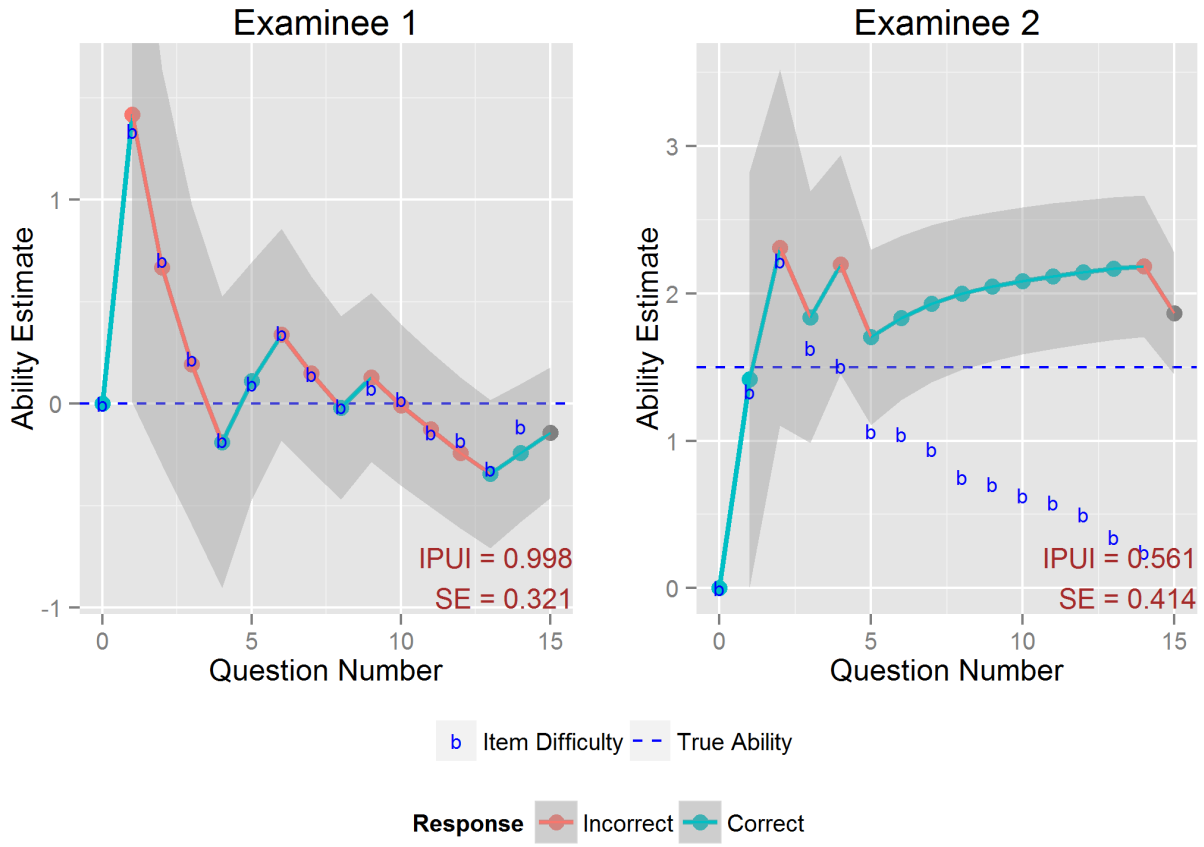

Figure 1.1: Adaptive Test Progress Plots for Two Examinees

The consequences of the inadequacy of an item pool can be serious depending on the stakes of the exam. For example, as seen in Figure 1.1, the standard error of the ability estimate for the second examinee is higher. The measurement error of examinees with similar abilities as Examinee 2 will be higher and this can affect the decisions made from their test scores. Also, even though Examinee 2 correctly answers each item, the items are getting

easier and easier. This is contrary to the basic premise of a CAT. Probably this will affect the motivation of the examinee. So, test developers should ensure that their item pools are supporting the purposes of the tests. The index developed in this dissertation will give test developers a tool to evaluate the quality of their item pools.

Test developers are very motivated to keep their item pools as efficient as possible. They want to use as many items as possible to ensure the quality of the test. But they don't want to use item pools that are more than enough. Because item pool development is an expensive enterprise. Breithaupt, Ariel, and Hare (2010) estimated that for a 40 item CAT that will span over 5 years with two administrations per year, 2000 items are necessary. Considering that developing one item with traditional item development methods (with all necessary control mechanisms) costs between $1,500-2,500 (Rudner, 2010), the total cost of an item pool reaches $3,000,000 to $5,000,000 (Gierl & Lai, 2013). As a result, test developers are motivated to reduce the size of their item pools and make their item pools as efficient as possible.

Evaluation of the item pool is very important because of the possible effects of a weak item pool on the decisions based on the test results. An inappropriate item pool can increase the biases and standard errors of the ability estimates. If the test is a variable length CAT, larger standard errors would increase the test length. If the item pool is sufficient for one group of examinees and not for another, this will impair the fairness of the test. Some group of examinees might have less precise ability estimates or longer tests depending on the test specifications. Simulations might reveal such problems. But tracing back to the source of such problems might not be straightforward, especially for CATs with complex designs.

Weak item pools might also cause the violation of some test specifications or administration of inappropriate items. Test developers should evaluate their item pools and ensure that their item pools are adequate for the test specifications of their CATs. For example, Eggen and Verschoor (2006) observed that the CAT algorithm they designed produced tests that did not meet the test specifications. The CAT algorithm failed to provide appropriate items to

the examinees. The authors attribute this to the lack of appropriate items in the item pool.

The motivation for this study is to create an index that quantifies the performance of an item pool for a given adaptive test and examinee population. The perfect item pool performance will be achieved when an item pool can provide a perfect item to an examinee regardless of the ability of the examinee or the stage of the test, while meeting all of the constraints of the test. A perfect item is an item with maximum possible amount of information at a given ability level. For example, for one-parameter logistic (1PL) model, the perfect item for a given ability has a difficulty parameter equal to this given ability value.

In practice, almost all item pools are imperfect. But evaluating the deficiency of an item pool is not straightforward. Adaptive tests are usually evaluated with outcome variables such as standard error of measurement (SEM), bias of the estimates, root mean squared error (RMSE), item exposure rates, overlap rates or decision accuracy. All of these are very valuable indicators to show the different aspects of the quality of an adaptive test. But the quality of an item pool cannot be evaluated solely by any of these indicators.

The quality of an item pool is intermingled with many aspects of an adaptive test such as the item selection procedures, the ability estimation mechanisms, constraints exposed on adaptive tests, and test specifications. When evaluating the outcomes of an adaptive test, usually it is not possible to single out the effects of each of these individual factors without performing a large simulation study. The index that is developed in this dissertation aims to quantify, for an item pool, the amount of deviation from a perfect item pool. A test developer will be able to use this index to evaluate the item pool's prospective performance. Consequently this will lead to a decision of either keeping the item pool intact, or improving it by adding more appropriate items, or removing the redundant items and saving them for future administrations.

CHAPTER 2

LITERATURE REVIEW

2.1 Notation

In this thesis, the notation used by van der Linden and Pashley (2010) has been followed. Items in the item pool are denoted by $i = 1, \ldots, I$. The order of presentation of items is denoted by $k = 1, \ldots, K$, where K is the test length. So, i_k denotes the index of the item in the item pool which is the kth selected item in the adaptive test. The set of items that are already administered before the selection of kth item is denoted as $S_{k-1} = \{i_1, \ldots, i_{k-1}\}$. The set of items remained in the item pool after the administration of $k-1$ items is denoted as $R_k = \{1, \ldots, I\} \setminus S_{k-1}$. The response string of an examinee will be denoted as $u_{i_1}, u_{i_2}, \ldots, u_{i_K}$. In this study only dichotomous items used, so the values of u_{i_k} can be either 0 for incorrect response or 1 for correct response. The ability parameter of examinees will be denoted by $\theta \in (-\infty, \infty)$. The examinees are indexed with $j \in \{1, \ldots, N\}$, where N is the total number of examinees.

2.2 Item Response Theory

Item response theory (IRT) is the backbone of a CAT. Almost all of the scoring and item selection algorithms use IRT. Especially in a CAT, since different examinees sees different items at different times, the existence of a common scale is paramount. IRT provides this common scale for items and examinees.

In IRT (Lord & Novick, 1968; Lord, 1980), the probability of a correct response of an

examinee with ability parameter θ to an item i is modeled as:

$$P\left(u_i = 1 \,|\theta\right) = c_i + (1 - c_i)\,\frac{e^{D \cdot a_i(\theta - b_i)}}{1 + e^{D \cdot a_i(\theta - b_i)}}$$
$$u_i = \begin{cases} 1 & \text{if item } i\text{'s response is correct} \\ 0 & \text{if item } i\text{'s response is incorrect} \end{cases} \tag{2.1}$$

where a is the item discrimination parameter, b is the item difficulty parameter, c is the lower asymptote parameter and D is the scaling factor which is usually taken as 1.7. The model in Equation (2.1) is called three-parameter logistic (3PL) model. Fixing the c parameter to 0 gives two-parameter logistic (2PL) model, and further fixing the a parameter to 1 gives 1PL model or Rasch model when $D = 1$ (Rasch, 1961).

2.3 Computerized Adaptive Testing

CAT is a complex method of delivering a tailored exam that adapts to an individual examinee. The research supporting the development of CAT has more than 40 years of history. It has been used in operational testing in the past twenty years.

Compared to paper and pencil (P&P) tests, CAT has several advantages such as shorter tests, increased test reliability, on demand testing, immediate test scoring and reporting (Meijer & Nering, 1999). A CAT uses half as many items compared to P&P tests (Weiss & McBride, 1984), in some cases even less (Gibbons et al., 2008). A CAT allows the measurement of information such as response times (Wise, Bhola, & Yang, 2006), speech entries, graphical entries, mouse movements and other tracking information that are not available to P&P tests. A CAT also allows the use of innovative items that can help to increase the validity evidence of tests which are not available in P&P tests (Luecht & Clauser, 2002).

The list below shows the basic algorithm of a CAT:

1. Specify an initial ability estimate ($\hat{\theta}_0$) as a starting point

2. Select an item from the available items in the item pool to deliver to the examinee

3. Score item and update examinee's ability estimate ($\hat{\theta}_k$)

4. Evaluate the termination criteria:

    a) If satisfied, conclude the test

    b) If not satisfied, go to step 2.

In the following sections each part of the list above will be explained further.

2.3.1 Initial Ability Estimate

In the first step of a CAT, an appropriate initial ability estimate should be designated. There are two options for the starting point of a CAT: variable starting point and fixed starting point. When a fixed starting point is used, each examinee is assigned the same initial ability estimate at the beginning of the CAT. In tests with variable starting point, some prior information about the examinee guides the choice of the initial ability estimate.

It is recommended that the initial ability estimate use all of the available information about the examinee (Kingsbury & Wise, 2000). Examples of this prior information might be the previous test results, school grades, student background information, and other relevant collateral information that is correlated with examinee's ability. Variable starting points will increase the efficiency of the adaptive test. For some examinees, this initial estimate might be off. But the adaptive nature of the test will solve this problem. Also, if the abilities of examinees are estimated using Bayesian estimation methods, the variable starting point will reduce the bias of these estimates (Weiss & McBride, 1984; Wang & Vispoel, 1998). For maximum likelihood estimation (MLE), Wang and Vispoel (1998) found negligible effects of using variable and fixed starting points on bias.

Using variable starting points has an added benefit to test developers too. If a fixed starting point is used for all examinees, each examinee will see the same items at the beginning

of the test (supposing that there is no exposure control). Consequently some items will be over exposed. Variable starting points can alleviate this problem.

Even though it's many psychometric benefits, variable starting points have an important drawback. Using information that possibly affect examinee's final ability estimate beyond the examinee's current test performance might be objectionable depending on the test purpose. This is the main reason why many high stakes tests use a fixed starting point.

Usually 0 is chosen as a fixed starting point for a CAT, because it is the middle of the ability distribution. But depending on the circumstances, different fixed starting points might be considered. For example, if the test length of a CAT is rather long and test developers want examinees to warm up to the exam and make an easy start, a lower starting point can be set. For licensure exams, the initial starting point might be set at the cut score of the test.

2.3.2 Item Selection

The item selection algorithm is the most important part of the adaptive test. At each stage, the CAT algorithm should select the most appropriate item usually with the existence of many constraints. Item information is an important determinant of the item selection processes. Almost all item selection algorithms try to select an item that will give the largest amount of information to the examinee. In this dissertation only a few of the available item selection algorithms were investigated in detail, but in the CAT literature there are many of them (van der Linden, 1998a; Barrada, Olea, Ponsoda, & Abad, 2010; van der Linden & Pashley, 2010).

2.3.2.1 Maximum Fisher Information

By far the most used item selection algorithm in a CAT is maximum Fisher information (MFI) (Lord, 1977a). In this method, the item that has the maximum amount of information at an examinee's intermediate ability estimate ($\hat{\theta}$) will be administered. For a test with K

items, the Fisher information function can be expressed as:

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\boldsymbol{u}|\theta)\middle|\theta\right] = \sum_{k=1}^{K}\frac{\left(P_{i_k}{}'\right)^2}{P_{i_k}Q_{i_k}} \tag{2.2}$$

where $L(\boldsymbol{u}|\theta)$ is the likelihood function defined as:

$$L(\boldsymbol{u}|\theta) = L\left(U_{i_1} = u_{i_1}, \ldots, U_{i_K} = u_{i_K}|\theta\right) = \prod_{k=1}^{K} P_{i_k}^{u_{i_k}}\left(1 - P_{i_k}\right)^{1-u_{i_k}} \tag{2.3}$$

where $P_{i_k} = P\left(u_{i_k} = 1|\theta\right)$ is defined in Equation (2.1). For the 3PL model substituting the values and calculating the derivative of Equation (2.1) for item i gives:

$$\mathcal{I}_i(\theta) = \frac{(Da_i)^2(1-c_i)}{\left(c_i + e^{D\cdot a_i(\theta-b_i)}\right)\left(1 + e^{-D\cdot a_i(\theta-b_i)}\right)^2} \tag{2.4}$$

At each step of the CAT, the MFI algorithm searches for an item that maximizes the total information (Equation (2.2)) given the previous S_{k-1} items.

MFI has some important advantages. At each step, it selects the most informative item which increases the efficiency of the CAT. The precision of the test increases rapidly as more items are administered. It is widely used and it's properties are well researched.

MFI method only uses the examinee's current test data to select a new item. This is desirable in some circumstances. But, usually at the early stages of the CAT, there is not enough information to guide this item selection algorithm. As a result, the items that are selected might not be the most appropriate ones. Additionally, the items at the beginning of the test cause big jumps in the ability estimates. This is the reason why test preparation companies tell their students to be extra careful while responding the first few items on the CAT. This problem might be alleviated by using prior information to select items or using different item selection paradigms such as Kullback-Leibler at least at the beginning of the test (Chang & Ying, 1996).

Because MFI method is selecting most informative items, it will reduce the uncertainty of the ability estimates, i.e. standard error (SE), more quickly. Han (2012) found that MFI

resulted in the lowest SE of estimate regardless of the exposure control. Even though this is desirable, one possible side effect of this might be on variable length CAT tests. For those tests, usually the test terminates when the SE drops below a predefined threshold. With MFI, there is a chance that tests terminate prematurely. Also, Warm (1989) showed that for shorter tests approximating the standard errors using information functions underestimate the true error variances.

MFI has some other disadvantages as well. For short tests, Chen, Ankenmann, and Chang (2000) found that MFI performed marginally worse than other item selection methods they investigated. For tests longer than 10 items, they found that this difference disappeared.

Using items solely based on their information values will result in disproportionate use of some highly informative items (Way, 1998). This has two disadvantages. First, highly informative items, which usually have high item discrimination parameters, exposed a lot. On the other hand, items with low item discrimination parameters might not be exposed at all. Second, since very informative items are used at the beginning of the test, where the provisional ability estimates are inaccurate, the final ability estimates will be over and under estimated (Chang, 2004). To mitigate this problem, methods such as a-stratified item selection method (Chang, Qian, & Ying, 2001) or a-stratified with b-blocking item selection method (Chang et al., 2001) has been proposed.

2.3.2.2 Owen's Bayesian Item Selection

Owen's Bayesian item selection algorithm (Owen, 1975) is based on the reduction of the posterior variances of the ability estimates. From a Bayesian framework, the posterior distribution of the ability given the responses to previous $k-1$ items is:

$$g(\theta|\boldsymbol{u}) = P\left(\theta \left| u_{i_1}, \ldots, u_{i_{k-1}} \right.\right) = \frac{P\left(\boldsymbol{u}, \theta\right)}{P\left(\boldsymbol{u}\right)} = \frac{L\left(\boldsymbol{u}\,|\theta\right) \cdot g\left(\theta\right)}{\int L\left(\boldsymbol{u}\,|\theta\right) \cdot g\left(\theta\right) d\theta} \tag{2.5}$$

where $g(\theta)$ is the prior distribution of the ability, which is usually a normal distribution. The calculation of this posterior distribution is not computationally simple. Owen used a normal approximation to this posterior distribution and he proved that as the number of administered items go to infinity, the expected value of the posterior distribution will converge to the true value of θ.

Each examinee will start the test with the initial ability estimate that is equal to the expected value of the prior distribution, $g(\theta)$. At each stage, Owen's item selection algorithm searches for an item that will reduce the posterior variance most. According to Owen this can be achieved by minimizing the β function (Vale & Weiss, 1977):

$$\beta_i = \frac{1}{(1-c_i)} \cdot \left(1 + \frac{1}{\sigma_0^2 a_i^2}\right) \cdot \left(1 + \frac{1}{K}\right) \cdot \left(c_i + \frac{1-c_i}{K}\right) \cdot e^{2D^2}$$

where

$$D = \frac{b_i - \mu_0}{\sqrt{2\left(a_i^{-2} + \sigma_0^2\right)}}$$

$$K^{-1} = \frac{1 - \mathrm{erf}(D)}{2}$$

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t.$$

where μ_0 and σ_0 are the mean and standard deviation of the prior distribution, respectively.

After the examinee answers an item, a new posterior distribution is computed using the item response and prior distribution. Then, this new posterior distribution becomes the prior distribution for the selection of the next item. When the variance of the posterior distribution reduces to an acceptable level that the test developer can tolerate, the test ends.

At each stage, posterior variance is calculated for each available item in the item pool. When the computer processing speeds were slow, this caused long waits after the examinee's response. Vale and Weiss (1977) proposed a rapid item search procedure to solve this problem at that time. But as the computer speeds increased this problem vanished. As discussed in Section 2.3.4.3, Owen's item selection algorithm is much faster than other Bayesian methods.

2.3.3 Constraints on Item Selection

In theory, the item that is most informative at the examinee's intermediate ability estimate should be administered, but in practice this rarely happens. There are several statistical or non-statistical rules that constrain the item selection. These constraints ensure that each test follows the test specifications as closely as possible and each examinee gets a comparable and standardized test. In addition they help test developers to secure their item pools.

In operational testing, the number of constraints on item selection can go up to 100. For example, van der Linden et al. (2006) reported that there are 96 constraints in LSAT test. In their adaptive test simulations, Stocking (1994) used up to 75 constraints on item selection.

The majority of the constraints on item selection are content balancing, exposure control and item enemies (Weiss, 2011). But constraints are not limited to these. Eignor, Stocking, Way, and Steffen (1993) gave a detailed account of such constraints. For example, a test developer might not want to use items that contain uncommon words more than once or twice in a test. Item format can be an important constraint as well. For instance, in a sentence completion section, test developer might want to preserve a certain ratio of items that contain a single blank as opposed to two blanks.

Some items should not be presented to an examinee within the same test. For example, item enemies which provides a clue to the solution of each other. Or redundant items that are very similar to each other. In such cases, the item selection algorithm should not select such items if one of them is already administered. In P&P tests, these items can be avoided before presenting the test to the examinees by checking the test forms. In a CAT, item enemies can be bundled in subsets. If an item within a subset is administered to an examinee, remaining items are removed from the available item pool for that examinee.

2.3.3.1 Content Balancing

Each test has an inference about examinee scores. If the inferences of a test are related to general mathematics ability and the majority of the test content is coming from trigonometry, then the test score do not reflect the claims of the test. For this reason, each test needs to follow a test blueprint. This blueprint delineates the details about the test including the content area distribution of items, cognitive requirements of items and etc. Content balancing is a mechanism needed for a CAT to follow the test specifications and to avoid over-testing or under-testing of some content areas. Most test specifications set desired content coverage ratios such that certain percentages of the test items come from each content domain.

Kingsbury and Zara (1989) proposed a simple and intuitive content balancing algorithm. First, test developer specifies the percentage of test items that should come from each content area, i.e. target percentages. After the administration of each item, the computer calculates the empirical percentages of each content area. Then, these empirical percentages are compared to the target percentages. The content area which has the largest discrepancy between the target and empirical percentages is selected. Items from other content areas are filtered out from the item pool and the next item will be delivered from the available items within this content domain. Basically in this method, item pool is partitioned into smaller item pools according to item content. At each stage an item from one of these smaller item pools is selected.

One problem with fixed number of items from each content is the potential interaction between item content and item difficulty (Segall, 1996). If, for example, trigonometry items are difficult items and arithmetic items are easier items in the item pool, a low ability examinee might need to answer trigonometry items that are way above his/her ability level. Clearly this reduces the efficiency of a CAT. A CAT using multidimensional item response theory (MIRT) might be more effective for such situations.

Besides this rather simple content balancing method, there are other techniques as well (He,

Diao, & Hauser, 2014). Examples of some other content balancing methods are the shadow test approach (van der Linden, 2010), weighted deviations model (Swanson & Stocking, 1993) and maximum priority index method (Cheng & Chang, 2009).

2.3.3.2 Exposure Control

Depending on the stakes of the examination, test developers may not want some items to be overused. As the stakes of the test increases, the incentive for test users to obtain the test items without permission increases. Some test preparation organizations even tried systematically to obtain test items (Davey & Nering, 2002). Therefore, test developers should protect their items from such attempts. Also test developers don't want some items to be underused due to the high cost of producing items. In order to control the usage of items in a CAT, the frequency of item administration is constrained using exposure control methods. Exposure control procedures are needed to maintain the fairness and the validity of the test by preventing examinees from having pre-knowledge of the items.

Test developers deal with the item exposure problems in two general ways. They deal with the item exposure problem before the examination by managing the item pools, and/or they deal with it during the examination by putting some constraints on the item selection algorithm.

Test developers may choose to use item pools for a certain amount of time and change it. The amount of time depends on the frequency and volume of the test administration and stakes of the test. Test developer can use a completely new item pool, a previous item pool or remove only the problematic and highly exposed items from the pool.

Item exposure can be controlled during the examination as well. Broadly, exposure control methods during the test administration can be divided into two (Way, 1998): methods based on randomization and methods based on the frequency of administration of items for a particular population.

There are many variations of the randomization approach to exposure control. The

randomesque procedure (Kingsbury & Zara, 1989) is a simple way of dealing with exposure control. Using this procedure, at each stage, the CAT algorithm randomly selects one item from the most informative m items where $m \in \{2, 3, \ldots\}$. Another method mentioned in Eignor et al. (1993) is a variation of this randomesque procedure. First item is selected from a group of eight best items, second item is selected from a group of seven best items and so on. After the eighth item the optimal item is selected. The idea behind this approach is, at the initial stages of the test, almost all examinees see the same set of items. After a certain number of items, there will be enough variation in the examinee responses to select an optimum item. Bergstrom, Lunz, and Gershon (1992) offered an exposure control method for a CAT using the Rasch model. In their method, an item with difficulty parameter within 0.10 logits of the intermediate ability estimate is selected randomly.

Sympson-Hetter exposure control (Hetter & Sympson, 1997) is an example of the second category of exposure control methods that are conditional on the frequency of administration of items for a particular target population. In this method, a maximum exposure rate between 0 and 1 is assigned to each item using simulations. Lower maximum exposure rates are assigned to items that are very informative and tend to be administered most. During the test, after selecting an optimum item for administration, a random number from a uniform distribution between 0 and 1 is generated and compared to the exposure rate of this item. If this random number is smaller than the maximum exposure rate of the selected optimum item, then this item is administered to the examinee. Otherwise, the next optimum item is selected and the same exposure control procedure is applied to this item as well until an item is administered to the examinee.

In addition to these two methods, there are many methods to control item exposure during the test (Revuelta & Ponsoda, 1998; Georgiadou, Triantafillou, & Economides, 2007). Even though it is very important to contain exposure rates at acceptable levels, these procedures will reduce the efficiency of a CAT.

2.3.4 Ability Estimation

After an examinee answers an item, the ability is estimated to either terminate the test or to select an item using this estimate. Selecting an appropriate estimation method is crucial. Estimation method will affect the final score that is reported, the decision to terminate the test and the items that are selected for administration. In the literature, there exists numerous ability estimation methods (Wang & Vispoel, 1998; van der Linden & Pashley, 2010). Three of these estimation methods are investigated further in this dissertation: MLE, expected a posteriori (EAP) and Owen's Bayesian ability estimation.

2.3.4.1 Maximum Likelihood Estimation

By far the most used method of ability estimation in a CAT is MLE. MLE depends on one of the main assumptions of IRT, local independence (Hambleton & Swaminathan, 1985). According to the local independence assumption, for any examinee the partial correlation between any two items will be zero when the ability parameter is held constant. As a result of this local independence assumption, the likelihood of a response string for a single examinee can be calculated using the following product:

$$L\left(U_{i_1} = u_{i_1}, \ldots, U_{i_K} = u_{i_K} | \theta\right) = \prod_{k=1}^{K} P_{i_k}^{u_{i_k}} \left(1 - P_{i_k}\right)^{1-u_{i_k}} \tag{2.6}$$

where P_{i_k} is defined as in Equation (2.1). MLE of the ability is the value of θ that maximizes this likelihood:

$$\hat{\theta}^{MLE} = \underset{\theta \in (-\infty, \infty)}{\arg\max} \left\{ L\left(\boldsymbol{u} | \theta\right) \right\} \tag{2.7}$$

Newton-Raphson method can be used to find the $\hat{\theta}$ value that maximizes Equation (2.6). In numerical analysis, the Newton-Raphson method is used to approximate the roots of a real-valued function (Hildebrand, 1987). In the case of Equation (2.6), the maximum value is the root of the first derivative of the likelihood function. The Newton-Raphson

procedure starts with an initial estimate for $x_0 = \theta_0$. Starting from this initial estimate, the Newton-Raphson procedure iteratively approximates to the root of the first derivative using the following equation:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.8}$$

where n is the iteration number and

$$f(x) = \frac{d}{d\theta} L(\boldsymbol{u}|\theta)$$
$$f'(x) = \frac{d^2}{d\theta^2} L(\boldsymbol{u}|\theta).$$

The iterations stop when the difference between x_{n+1} and x_n becomes acceptably small. The user defines this acceptable small value. At the end, the process said to be converged and the final x_{n+1} value is the MLE of the θ, denoted as $\hat{\theta}$. Some authors (Hambleton & Swaminathan, 1985) choose to maximize the natural logarithm of likelihood function ($\ln(L(\boldsymbol{u}|\theta))$) instead of likelihood function. Both methods will converge to the same number. But maximizing the logarithm of likelihood is preferable to maximizing the likelihood, because logarithm of likelihood reduces to the summation of the probabilities and the numbers used in the analysis is less extreme when the sums rather than products are used.

The Newton-Raphson algorithm is a quick method to find the ability estimate of a given response string. It is important to select a good initial estimate for this approximation to be quick. In some cases there might be some local minimum or maximum values. The algorithm might converge to these values instead of a global maximum point. The user should be aware of such possibilities and choose good starting values to converge to the global maximum value.

MLEs have many desirable properties (Hambleton & Swaminathan, 1985). They are consistent, as the number of items increases the estimates converge to their true values. They are efficient, they have the smallest variance asymptotically. And asymptotically they are

normally distributed. The last property is very useful in practice. It allows the calculation of the standard error of the maximum likelihood estimator:

$$se(\hat{\theta}|\theta) = \frac{1}{I(\hat{\theta})^2} \tag{2.9}$$

where $I(\hat{\theta})$ is the information function given in Equation (2.2).

On the other hand, MLE has an important practical disadvantage. For response strings that consist of all 1's or all 0's it will tend to ∞ or $-\infty$, respectively. To avoid this problem, the CAT test can start with a Bayesian estimation procedure and switch to MLE after obtaining a heterogeneous response string.

2.3.4.2 Expected a Posteriori Estimation

In EAP estimation the information from examinee's responses are combined with the information about the population (Bock & Mislevy, 1982). EAP is the expected value of the posterior distribution in Equation (2.5) which can be written as:

$$\hat{\theta}^{EAP} = E(\theta|\boldsymbol{u}) = \int_{-\infty}^{\infty} \theta g(\theta|\boldsymbol{u}) d\theta \tag{2.10}$$

Variance of the EAP estimate can be written as:

$$\text{var}(\theta|\boldsymbol{u}) = \int_{-\infty}^{\infty} \theta^2 g(\theta|\boldsymbol{u}) d\theta - [E(\theta|\boldsymbol{u})]^2 \tag{2.11}$$

As mentioned in Section 2.3.2.2, the calculation of these integrals are not trivial. A common approach is to approximate the values of these integrals using numerical integration methods.

Another estimation method that uses the same posterior distribution is maximum a posteriori (MAP) estimation (Samajima, 1969). Instead of finding the expectation of posterior distribution like EAP, MAP locates the mode of the posterior distribution:

$$\hat{\theta}^{MAP} = \underset{\theta \in (-\infty, \infty)}{\arg\max} \{g(\theta|\boldsymbol{u})\} \tag{2.12}$$

2.3.4.3 Owen's Bayesian Estimation

Owen's Bayesian estimation (Owen, 1969, 1975) is a sequential ability estimation method. It eliminated the burdensome computations of MLE. At each step of a CAT, the posterior distribution of the ability from the previous step is used as a prior distribution for the estimation of ability. Furthermore, assuming a normal distribution as a prior distribution for examinee population enables a closed form approximation to the posterior mean and variance of the ability:

$$\begin{aligned} M_k &= M_{k-1} - \frac{V_{k-1}}{\sqrt{a_k^{-2} + V_{k-1}}} \cdot \frac{\phi\left(D_k\right)}{\Phi\left(D_k\right)} \cdot \left(1 - \frac{u_k}{A_k}\right) \\ V_k &= V_{k-1} - \frac{V_{k-1}^2}{\sqrt{a_k^{-2} + V_{k-1}}} \cdot \lambda_k \end{aligned} \tag{2.13}$$

where M_k and V_k are the mean and variance of the posterior distribution, M_{k-1} and V_{k-1} are the mean and variance of the prior distribution, a_k, b_k, c_k are the item parameters of the kth item, u_k is the item response. D_k, A_k and λ_k in Equation (2.13) are

$$D_k = \frac{b_k - M_{k-1}}{\sqrt{a_k^{-2} + V_{k-1}}}$$

$$A_k = c_k + (1 - c_k) \cdot \Phi\left(-D_k\right)$$

$$\lambda_k = \frac{\phi\left(D_k\right)}{\Phi\left(D_k\right)} \cdot \left(1 - \frac{u_k}{A_k}\right) \cdot \left[\left(1 - \frac{u_k}{A_k}\right) \cdot \frac{\phi\left(D_k\right)}{\Phi\left(D_k\right)} + D_k\right]$$

where ϕ and Φ are the probability density function and cumulative density function of the standard normal distribution. Mean of the posterior distribution in Equation (2.13) corresponds to the ability estimate and square root of the variance of the posterior distribution corresponds to the standard error of the ability estimate.

Owen's Bayesian estimation was very popular due to it's computational simplicity due to its closed form equations. But it has a major disadvantage. At each step, it uses a normal density function to create a posterior distribution (Wang & Vispoel, 1998), but in fact the shapes of the distributions can deviate from the shape of normal distributions. As a result,

this estimation method introduces a bias to the estimates. For example, simply changing the order of the administration of items might change the ability estimate.

MLE is generally preferable to Bayesian estimation methods. In the long run, MLE is asymptotically unbiased. It is not affected by any other factor, like Bayesian estimation methods, other than actual test performance. Bayesian estimation methods yield biased estimates but the standard errors associated with them are smaller (Lord, 1986; Wang & Vispoel, 1998).

2.3.5 Item Pools in CAT

Parshall et al. (2002) defined an item pool as a "collection of test items that can be used to assemble or construct a computer-based test" (p. 21). In the literature, item pools have been called as "item banks", "question banks", "item collections", "item reservoirs" and "test item libraries" (Millman & Arter, 1984). Even though there might be subtle differences between these terms, they are used synonymously in this study.

The quality of the item pool is very crucial because the quality of the CAT depends on them. Chapter 1 explained the importance of the item pools for adaptive tests. According to Flaugher (2000), a satisfactory item pool for adaptive testing should have items with three characteristics: (1) high item discriminations ($a > 1$) (2) a rectangular distribution of item difficulties and (3) low guessing parameters ($c < 0.2$). McBride (1977) described an ideal item pool as having a large number of highly discriminating items ($a > 0.8$) and a rectangular distribution of item difficulties that covers the ability continuum. A similar definition of ideal item pool was given by Mills and Stocking (1996). Urry (1977) gave a more detailed description of item parameters of an item pool: item discriminations should be larger than 0.8, item difficulty parameters should be evenly and widely distributed between -2 and 2, item guessing parameters should be lower than 0.3. Urry added that the item pools should have at least 100 items.

The quality of the items in the item pool is important because the examinees are

administered relatively fewer items compared to P&P tests. A flawed item in an item pool has many perils. It affects the ability estimates, which consequently affects the subsequent items administered. It has a larger effect on the ability estimates because fewer items are administered in CATs. Since the items each examinee sees are different, a flawed item can affect some examinees but not others, which in turn hampers the fairness of the test (Wainer, 2000).

2.3.5.1 Item Pool Size

The size of the item pool is another consideration for the quality of an item pool. The size and the item difficulty distribution of an item pool depends on the CAT specifications and the examinee population (Reckase, 2010). Stocking (1994) listed six factors that affect the size of an item pool: "item selection algorithm, constraints on item content, psychometrics, and exposure, stopping rules, overlap restrictions, test scoring, requirements of parallelism with existing paper-and-pencil forms" (p. 7).

The purpose of the test can affect the size of an item pool (Parshall, 2002). Larger item pools are needed for high stakes tests compared to low stakes tests. In high stakes tests, due to the stakes of the decision, test scores should be more precise. High stakes tests are more prone to cheating which requires more limits over the exposure of items. Another consideration for item pool size is the number of test days (i.e. length of the testing window). As the number of test dates increased, the items in the pool exposed more. This creates a need for more items in the item pool (Parshall, 2002).

In the literature there is not a consensus regarding the size of an item pool. Urry (1977) advised an item pool with at least 100 items for a CAT test that improves the accuracy compared to a similar P&P test. Stocking (1994) recommended that a CAT item pool should be 12 times as large as the length of the CAT test. Chen, Ankenmann, and Spray (2003) advised that an item pool should be at least 6.7 times larger than a fixed length tests to contain the overlap rate between examinees below 15%. For an overlap rate below 10%, item

pool size should be 10 times as large as a fixed length test.

Very large item pools are not advisable in practice. Such item pools are difficult to manage, they can strain the hardware and software running the CAT test (Mills & Stocking, 1996). Searching a small item pool is faster compared to a large item pool. This is an important practical constraint for operational adaptive tests (Vale & Weiss, 1977). A single breach to the item pool might compromise a lot of items at the same time. In this sense, creating item pools with sufficiently enough items is important.

2.3.5.2 Item Pool Design and Assembly

Item pools for a CAT can be assembled in a similar fashion to the traditional tests. Initially, test developer defines a goal for the test. This goal might be measuring every examinee as precisely as possible, measuring high ability examinees precisely for a scholarship, making a decision at a cut point to obtain a license for a job, etc. The test developer then develops an item pool information function for this goal. Finally, the test developer assembles items so that the item pool information function matches the target information function. van der Linden (1998b) discussed several methods to assemble regular P&P tests using these three steps. These test assembly methods can be extended to the item pools of the CATs.

Some testing agencies have large item banks that contain many items. These large item banks are called the "master pool" or "vat" (Way, 1998). At each testing window, a test agency assembles an item pool from this vat that meets the test specifications. The item pools are replaced with the new ones after some conditions met. Way, Steffen, and Anderson (2002) calls them the docking rules. Item pools can be renewed after a certain number of examinees sees the item pool, or after a certain period of time. Item pools are assembled from the master item pool using the assembly techniques described in the previous paragraph. In addition, some researchers proposed methods to design optimal blueprint for an item pool and selecting items from the master pool using these optimum blueprints (Veldkamp & van der Linden, 2010). Reckase (2010) developed the bin-and-union method to build an

optimum item pool blue print. This method was used in this study to build optimum item pools (Section 4.2.2.1). See He and Reckase (2013) for an example use of this approach.

2.3.6 Evaluation of the Item Pools

In the CAT literature there is not a specific method to evaluate the quality of item pools. Generally, item pools are compared performing simulations that use different item pools while holding everything else the same (Thompson & Weiss, 2011). The test developer checks the indicators like bias, SE, mean squared error (MSE), exposure rates and overlap rates and makes a decision about the quality of the item pool. Even though these are very valuable indicators, none of these are the direct measures of the quality of an item pool. For example, the reason why SE is not a direct measure of item pool quality is discussed throughly in Section 3.4 on page 31.

The most common method to evaluate item pool quality is investigating the item pool information function (Segall, Moreno, & Hetter, 1997). Item pool information functions of different item pools are compared while keeping the testing purpose in mind. For example, Xing and Hambleton (2004) compared six item pools using this method (Figure 2.1). Depending on the purposes of the test, test developer builds a target information function for the item pool. The item pools that are close to this target of information curve will be selected.

Another method of item pool evaluation is to investigate the exposure and overlap rates of the items within the item pools (Chang, 2004). An item pool that has a lot of overused or underused items is deemed to be an inefficient item pool. Also, for security reasons, test developers do not want high item overlap rates between two randomly selected examinees. These methods are important to monitor the quality of the item pool but it does not say much about the quality of the items an examinee sees. A test with high exposure rates might present high quality items to examinees, or an item pool with low exposure rates might not provide appropriate items to the examinees. A test developer might implement an exposure

[1]This graph is taken from Xing and Hambleton (2004, p. 8)

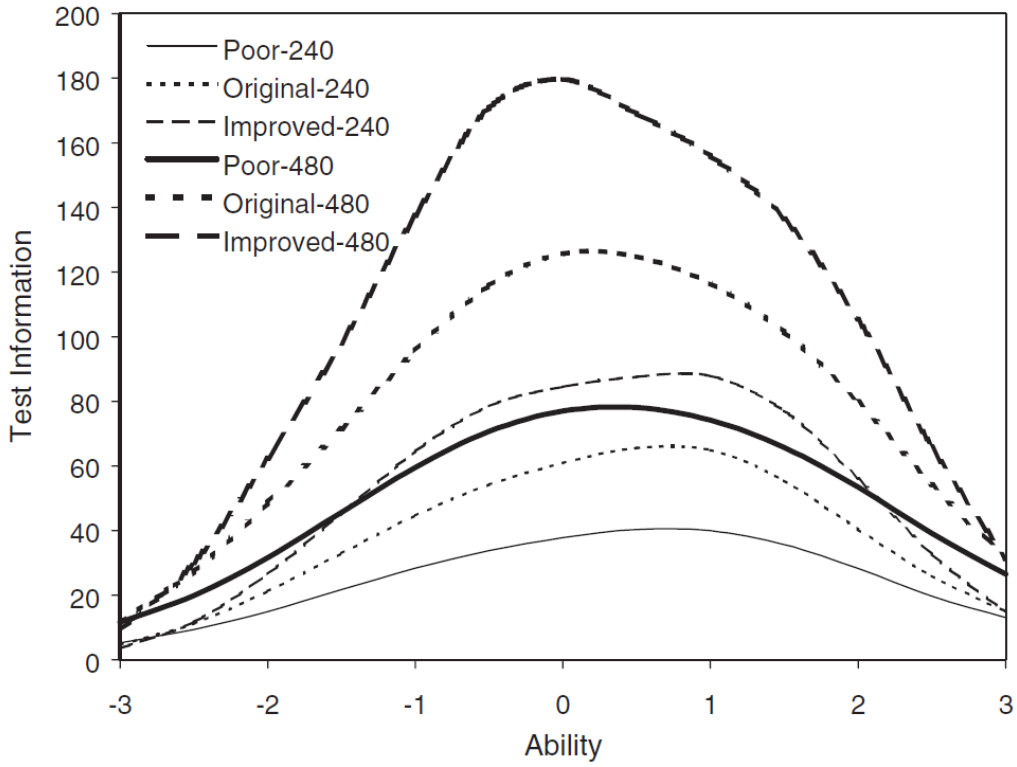

Figure 2.1: Comparison of the Information Functions of Six Item Pools Using the Item Pool Information Functions[1]

control to reduce the exposure rates of the items within the item pool. But this may result a loss in the efficiency of the CAT test.

CHAPTER 3

THE ITEM POOL UTILIZATION INDEX

3.1 Relative Efficiency

The origins of item pool utilization index goes back to the concept of relative efficiency. In Lord and Novick (1968), Birnbaum introduced the concept of relative efficiency. He used relative efficiency to compare two scoring methods. Lord (1980, p. 83) defined relative efficiency of two tests, x and y, at a certain ability θ as:

$$RE\left(y,x\right)=\frac{\mathcal{I}\left\{\theta,y\right\}}{\mathcal{I}\left\{\theta,x\right\}} \tag{3.1}$$

where $\mathcal{I}\left\{\theta,y\right\}$ and $\mathcal{I}\left\{\theta,x\right\}$ are the information functions for tests y and x, respectively. Lord (1974, 1975, 1977b, 1980) demonstrated the use of relative efficiency to evaluate and compare different tests.

As an example for the use of relative efficiency in comparing two tests, let's consider two tests with 10 items. Item parameters of these tests are in Table 3.1. If the test information function (TIF) of these two tests (Figure 3.1) are investigated, it can be seen that these two tests have different characteristics. Test 2 (green line) provides more information for examinees just above $\theta = 0$, but it is less informative for examinees at the extremes. On the other hand, Test 1 (red line) is not giving as much information as Test 2 for examinees with θ's between -0.5 and 1.5, but throughout the ability scale it provides more information. Relative efficiency of Test 1 to Test 2 (blue line) shows this. Test 1 is more efficient compared to Test 2 when blue line is above the dashed line $x = 1$. When blue line falls below the dashed line, Test 2 is more informative.

	a	b	c
Item1	1.21	-1.41	0.22
Item2	1.32	-1.21	0.14
Item3	1.16	0.01	0.17
Item4	1.57	0.25	0.23
Item5	1.32	-0.82	0.15
Item6	1.41	0.24	0.26
Item7	1.16	-0.55	0.15
Item8	1.12	0.51	0.18
Item9	1.35	-1.22	0.15
Item10	1.24	2.23	0.21

(a) Test 1

	a	b	c
Item1	1.08	-0.98	0.20
Item2	1.36	-0.70	0.15
Item3	0.97	0.37	0.22
Item4	1.29	-0.07	0.14
Item5	1.46	0.06	0.19
Item6	1.48	0.39	0.18
Item7	1.41	-0.75	0.18
Item8	0.69	0.55	0.19
Item9	1.07	0.17	0.23
Item10	1.56	0.36	0.14

(b) Test 2

Table 3.1: Item Parameters of Test 1 and Test 2

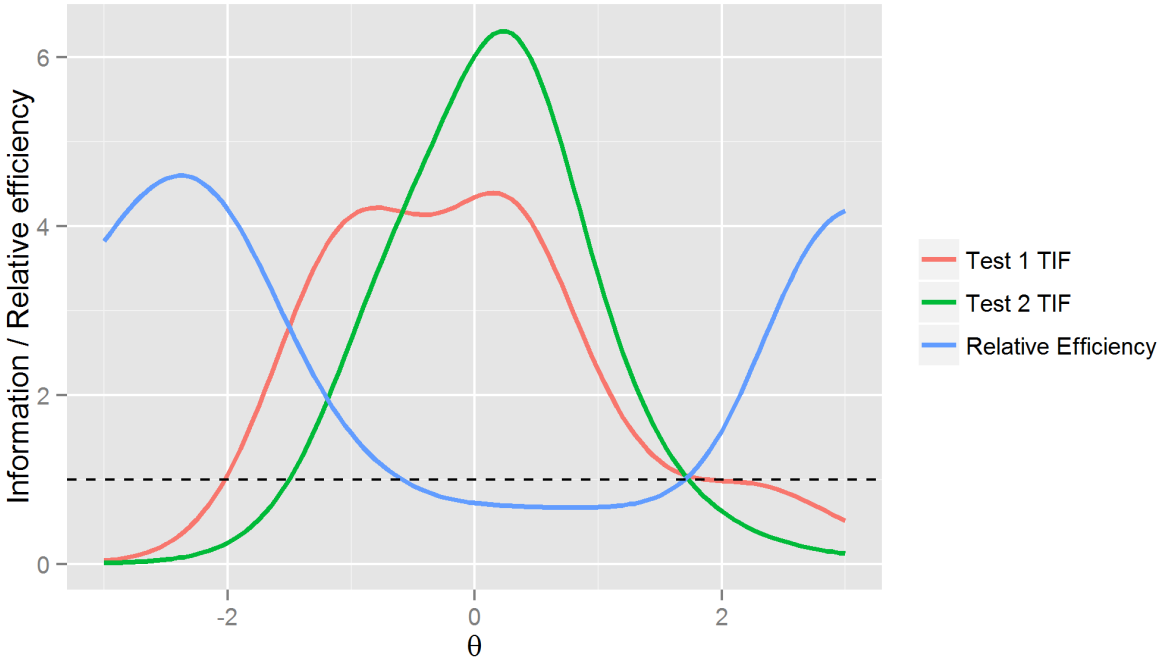

Figure 3.1: Test Information Functions and Relative Efficiencies of Test 1 and Test 2

More recently Han (2012) created an item selection algorithm for CAT using relative efficiency concept. In this algorithm he utilized from the concept of expected item efficiency, which is the "level of realization of an item's potential information at interim $\hat{\theta}$" (Han, 2012, p. 227). Mathematically an item i's expected efficiency after the administration of kth item

defined as:

$$\frac{\mathcal{I}_i\left[\hat{\theta}_k\right]}{\mathcal{I}_i\left[\theta_i^*\right]} \tag{3.2}$$

where $\mathcal{I}_i\left[\hat{\theta}_k\right]$ is the amount of information item i has at the interim ability estimate $\hat{\theta}_k$, and $\mathcal{I}_i\left[\theta_i^*\right]$ is the maximum potential information item i can have. For 1PL and 2PL models (Hambleton & Swaminathan, 1985), item i reaches maximum information at $\theta_i^* = b_i$, where b_i is the item difficulty parameter. For 3PL model, item i reaches maximum information at

$$\theta_i^* = b_i + \frac{1}{Da_i}\ln\left(\frac{1+\sqrt{1+8c_i}}{2}\right). \tag{3.3}$$

3.2 Item Pool Utilization Index

Han (2012) used expected item efficiency as an intermediate step to select most appropriate item for an examinee. He did not mentioned it as a possible way of evaluating an item pool's efficiency. This is where current research diverts from his research and the rest of the literature. In this dissertation, a slightly different version of item efficiency will be used to evaluate an item pool's performance. In the paper of Han (2012), the focus was whether an item's maximum potential is fulfilled or not (Equation (3.2)). In this dissertation, the focus is whether a perfect item is administered to an examinee or not. Even though these two interpretations give same results for basic IRT models, they are conceptually different. In this dissertation, the item efficiency for item i at $\hat{\theta}$ defined as:

$$\frac{\mathcal{I}_i\left[\hat{\theta}\right]}{\mathcal{I}_{max}\left[\hat{\theta}\right]} \tag{3.4}$$

where $\mathcal{I}_i[\hat{\theta}]$ is the information of item i at $\hat{\theta}$, and $\mathcal{I}_{max}[\hat{\theta}]$ is the value of information at $\hat{\theta}$ if an optimum item is administered to the examinee with $\hat{\theta}$. If this item efficiency is calculated for each item in an adaptive test for an examinee and then averaged, we will get:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{\mathcal{I}_{i_k}\left[\hat{\theta}_{k-1}\right]}{\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right]} \tag{3.5}$$

where K is the test length, i_k is the index of kth administered item in the item pool, $\hat{\theta}_{k-1}$ is the intermediate ability estimate after the administration of the $(k-1)$th item, $\mathcal{I}_{i_k}[\hat{\theta}_{k-1}]$ is the amount of information item i_k has at $\hat{\theta}_{k-1}$ and $\mathcal{I}_{max}[\hat{\theta}_{k-1}]$ is the amount of maximum information an optimum item has at $\hat{\theta}_{k-1}$. For $k = 1$, $\hat{\theta}_{k-1}$ becomes $\hat{\theta}_0$ which is the initial ability estimate. The value in Equation (3.5) reflects to what degree the items presented to an examinee deviate from the items coming from a perfect item pool. Here, a perfect item pool is defined as an item pool in which, whenever a CAT algorithm searches for an item to deliver, item pool can present a perfect item for that ability level, regardless of the stage of the test. A perfect item for a particular θ is defined as an item that provides maximum possible information at that θ level. This approach is similar to Eignor et al. (1993):

> ...an item is considered to have optimum statistical properties if it is most informative at an examinee's current maximum-likelihood estimate of ability. (p. 10)

This conceptualization of a perfect item is purely statistical and from the framework of IRT. Clearly, a perfect item and item pool should be valid for the intended purpose and use of the test scores (Kane, 2013). Another assumption in this definition of perfect item is that, a perfect item for an examinee is the item that is equal to the current ability estimate of the examinee. This might not be the case in every situation. Depending on the purpose of the test, a better item might be the one that maximizes the information at another ability level. In a licensure examination, for example, test developer might want to administer items that have maximum information at the cut score. The aim of the licensure examination is to learn whether examinees are above or below the cut score. The exact locations of the examinees might not be the primary aim of the examination. In this case, a perfect item might be defined as an item that has the maximum amount of information at the cut score.

Aggregating Equation (3.5) over a representative group of examinees will give information about how an item pool performs for that examinee group. The item pool utilization index

(IPUI) proposed for evaluating the performance of an item pool is:

$$IPUI = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\mathcal{I}_{i_{jk}}\left[\hat{\theta}_{j(k-1)}\right]}{\mathcal{I}_{max}\left[\hat{\theta}_{j(k-1)}\right]}\right) \tag{3.6}$$

where N is the total number of examinees that took the adaptive test, K_j is the test length of examinee j, i_{jk} is the index of kth test item presented to jth examinee within the item pool and $\hat{\theta}_{j(k-1)}$ is the ability estimate of examinee j after the administration of kth item. The first summation is taken over the examinees taking the adaptive test, and the second inner summation is taken over the specific test of each examinee.

The values that IPUI can take ranges between 0 and 1. An IPUI value of 1 signifies a perfect item pool. An IPUI value of 0 is theoretically possible. If each item presented to examinee has 0 information value at the examinee's current ability estimate, then IPUI will take a value of 0. But in practice, each item provides some information about an examinees. So, IPUI can get very close to 0 but cannot be 0 in practical settings.

IPUI can tell an item pools' level of efficiency. If one adds redundant items to an item pool that cannot be utilized by a CAT algorithm, IPUI will not increase or increase minimally. In this sense, IPUI is an indicator of the item pools' deficiency, instead of redundancy. IPUI can be used to diagnose an item pool. A test developer can calculate IPUI for certain groups of examinees or conditional on different ability levels and observe the weak spots of the item pool.

IPUI can be helpful to test developers and test users at two levels. First, IPUI can be used at examinee level as shown in Equation (3.5). An IPUI can be calculated for each examinee and both test developers and test users can monitor the quality of the item pool at this level. A test developer can assign a baseline IPUI value for each examinee and ensure that item pool is sufficient for each individual examinee. This will substantiate the fairness claim of the test. Second, as shown in Equation (3.6), IPUI can be used at the aggregate level. This allows test developer to get an overall picture of an item pool's performance at the group level. Aggregating IPUI at group level allows developer to weight the IPUI for

a target group. This might be necessary in cases where the test has a particular aim and target population. Test developer might want to see how an item pool performs for most of the target population and give a smaller weight to examinees at the extremes. Aggregation of the IPUI can be useful at such situations.

In contrast to the other outcomes of the CAT such as SE, MSE or bias, IPUI is a standardized measure that ranges between 0 and 1. It is not straightforward to compare two different values of SE because they depend on the context of the measurement. The ranges of other outcomes of a CAT are unspecified. Theoretically, SE and MSE can take any positive value, bias can take any value. For example, it is difficult to interpret whether a SE value of 0.4 is large or small without knowing the context. On the other hand, an IPUI value close to 1 always indicates an adequate item pool. IPUI values can be compared across different testing scenarios because they are dimensionless.

3.3 An Example Calculation of IPUI

The calculation of IPUI is very straightforward. An example IPUI calculation for Examinee 2 in Figure 1.1 demonstrated in Table 3.2.

$IPUI_k$ column shows the relative quality of the selected item at each step of the CAT. As can be seen from the first line, the initial ability estimate is $\hat{\theta}_0 = 0$. The difficulty parameter of the selected item ($b_{i_1} = 0.0013$) is almost equal to this initial ability estimate. The information value of this item ($\mathcal{I}_{i_1}[\hat{\theta}_0] = 0.722$) at this initial estimate also reached the maximum possible information value ($\mathcal{I}_{max}$). Examinee 2 gives a correct answer to the first item and the ability of the examinee is updated to $\hat{\theta}_1 = 1.42$. The item that can provide highest information at this ability estimate has an item difficulty parameter $b_{i_2} = 1.34$. This item is also very informative ($\mathcal{I}_{i_2}[\hat{\theta}_1] = 0.719$), but not as informative as the first item. IPUI value for this item slightly dropped to 0.995. After the third item, the discrepancy between examinee's estimated ability and the difficulty parameter of the selected item start to increase.

k	$\hat{\theta}_{k-1}$	b_{i_k}	u_k	$\hat{\theta}_k$	$\mathcal{I}_{i_k}\left[\hat{\theta}_{k-1}\right]$	$\mathcal{I}_{max}$	$IPUI_k$	$IPUI_{1:k}$
1	0.000000	0.001347672	1	1.417598	0.722	0.722	1.000	1.000
2	1.417598	1.338505825	1	2.311825	0.719	0.722	0.995	0.998
3	2.311825	2.226290085	0	1.839344	0.719	0.722	0.995	0.997
4	1.839344	1.631980644	1	2.198057	0.701	0.722	0.970	0.990
5	2.198057	1.510722847	0	1.703028	0.523	0.722	0.724	0.937
6	1.703028	1.068875300	1	1.833632	0.547	0.722	0.758	0.907
7	1.833632	1.046853738	1	1.931229	0.476	0.722	0.659	0.871
8	1.931229	0.946678729	1	2.001291	0.384	0.722	0.532	0.829
9	2.001291	0.754537673	1	2.047349	0.277	0.722	0.383	0.779
10	2.047349	0.707118514	1	2.086142	0.244	0.722	0.337	0.735
11	2.086142	0.630249407	1	2.117851	0.207	0.722	0.286	0.694
12	2.117851	0.584525037	1	2.145407	0.185	0.722	0.256	0.658
13	2.145407	0.500193537	1	2.168164	0.157	0.722	0.217	0.624
14	2.168164	0.347522088	1	2.185140	0.120	0.722	0.166	0.591
15	2.185140	0.247744310	0	1.864599	0.100	0.722	0.138	0.561

Table 3.2: IPUI Calculation Example

At the last item, even though the examinee's estimated ability based on his/her previous 14 responses is $\hat{\theta}_{14} = 2.19$, item pool can only provide an item with difficulty parameter $b_{i_{15}} = 0.25$. The amount of information that this item provides at this ability estimate is very low compared to previous items ($\mathcal{I}_{i_{15}}[\hat{\theta}_{14}] = 0.1$). This discrepancy reflected on the IPUI value for this item ($IPUI_{15} = 0.138$). At the end of the test, the overall value of IPUI for Examinee 2 is 0.561. Compare this value to the IPUI value of Examinee 1, 0.998.

3.4 Difference between IPUI and Standard Error

The biggest difference between IPUI and the SE is the θ values used to calculate the information function. When calculating the SE, the information formula will use only the final θ estimate (Equation (2.4)). In this sense SE is blind to the quality (or appropriateness) of the items which are presented at the intermediate stages of the adaptive test. SE only cares whether good items are presented to examinee which are around examinee's final ability estimate. If an examinee starts a CAT with a low initial ability estimate and improves her ability estimate continuously by correctly answering the items presented, SE will indicate

that the quality of the test is low because examinee get a lot of inappropriate items close to her final ability estimate. SE provides a very important piece of information. But it says little about the appropriateness of the items that were presented to the examinee. IPUI, on the contrary, can exclusively give information regarding the quality of the items presented.

The distinction between SE and IPUI is demonstrated with an example. Figure 3.2 shows the CAT processes of two examinees that are taking the same adaptive test with same test specifications and item pool.

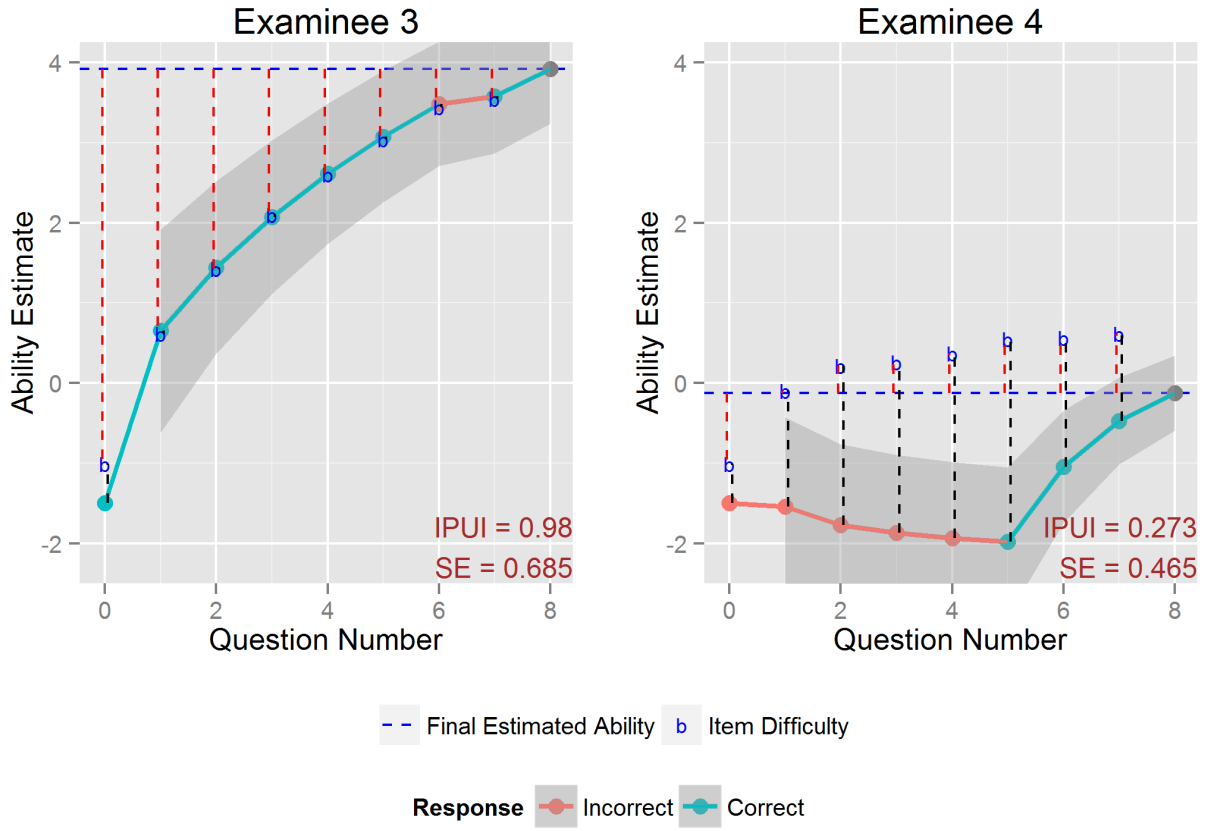

Figure 3.2: A Demonstration of the Difference between SE and IPUI

Examinee 3 is a high ability examinee and answered most of the questions correctly. Also, the item pool was very appropriate for Examinee 3. At each step, there was an appropriate item to present in the item pool which was very close to her intermediate ability estimate. Consequently, the IPUI value is very high for this examinee, 0.98. On the other hand,

Examinee 4 is an average ability examinee. She started test with consecutive incorrect responses. Later on the test, her responses improved. Since the item pool is composed of rather difficult items, Examinee 4 could not get appropriate items. Even though, she responded incorrectly, the difficulty of the items kept increasing. This result in a low IPUI value, 0.273.

When the SEs of these two examinees are examined, it is observed that Examinee 3 has a higher SE. As explained in the first paragraph of this section, SE calculates the information function with respect to the final ability estimate (red dashed lines in Figure 3.2). For Examinee 3, the difficulty parameters of the administered items were away from the final ability estimate. This produced in a high SE value for this examinee. For Examinee 4, even though the item pool was not very appropriate, the difficulty parameters of the items presented to this examinee were close to the final ability estimate. This produced a lower SE for this examinee.

Visually the difference between SE and IPUI can be conceptualized with the help of red and black dashed lines in Figure 3.2. SE is aggregating the distances shown by the red dashed lines. The longer the red lines, the higher the SE will be. IPUI aggregates the distances shown by the black dashed lines. The shorter the distances, the higher the IPUI will be.

In this example, SE clearly says nothing about the quality of the item pool. But, IPUI shows the deficiency of the item pool. The example shown above is not common in practice. But it shows the difference between SE and IPUI plainly. For most of the cases, SE and IPUI will be highly correlated. Tests with better item pools will have on average lower standard errors.

3.5 The Limitations of IPUI

In Equation (3.5), $\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right]$ is a equal for all items for 1PL model. For 1PL model, an optimum item which has maximum information at $\hat{\theta}_{k-1}$ has item difficulty parameter

$b_{i_k} = \hat{\theta}_{k-1}$. The maximum value of the information of this item is:

$$\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right] = \frac{D^2}{e^{D\cdot\left(\hat{\theta}_{k-1}-b_{i_k}\right)}\left(1+e^{-D\cdot\left(\hat{\theta}_{k-1}-b_{i_k}\right)}\right)^2} = \frac{D^2}{4} \tag{3.7}$$

For the 2PL model, an optimum item which has maximum information at $\hat{\theta}_{k-1}$ also has item difficulty parameter $b_{i_k} = \hat{\theta}_{k-1}$. Consequently, the maximum value of information for the 2PL model will be:

$$\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right] = \frac{D^2 a_{i_k}^2}{e^{D\cdot a_{i_k}\left(\hat{\theta}_{k-1}-b_{i_k}\right)}\left(1+e^{-D\cdot a_{i_k}\left(\hat{\theta}_{k-1}-b_{i_k}\right)}\right)^2} = \frac{D^2 a_{i_k}^2}{4} \tag{3.8}$$

For the 3PL, from Equation (3.3), the maximum value of information will be:

$$\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right] = \frac{\left(Da_{i_k}\right)^2\left(1-c_{i_k}\right)}{\left(c_{i_k}+\frac{1+\sqrt{1+8c_{i_k}}}{2}\right)\left(1+\frac{2}{1+\sqrt{1+8c_{i_k}}}\right)^2} \tag{3.9}$$

It can be easily seen from Equation (3.7) that the value of $\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right]$ does not depend on either item or ability parameters for 1PL model. The value is constant for all items. The maximum information for a perfect item can be captured by a constant. On the other hand, in Equations (3.8) and (3.9) the value of $\mathcal{I}_{max}\left[\hat{\theta}_{k-1}\right]$ depends on the values chosen for a_{i_k} and c_{i_k} parameters. Since we defined the perfect item as the one that maximizes the item information value, for the 2PL model, Equation (3.8) is maximized when $a_{i_k} = \infty$ and $b_{i_k} = \hat{\theta}_{k-1}$. For the 3PL model, the maximum information value will be reached when $a_{i_k} = \infty$, $c_{i_k} = 0$ and $b_{i_k} = \hat{\theta}_{k-1}$. The maximized values of informations for the 2PL and 3PL models will be infinite. This will pose a problem for the IPUI. The denominator of the Equation (3.6) will be infinite, which will force IPUI to be 0 for any practical testing situation using 2PL or 3PL model. This problem is acknowledged by Reckase (2010) where he discusses the impossibility of developing optimal item pools for the 2PL model.

One possible solution to this problem is fixing the value of the a parameter to a high value that is rarely reached, like 3. But any value chosen in this manner will be arbitrary. Another option might be to ignore the a parameter when calculating IPUI. This will bypass the infinity problem stated above but also loses valuable information about the quality of item pool. An item's quality is highly related to it's discrimination power. Ignoring this will result an equal IPUI value for a quality item pool with many highly discriminating items and an item pool with many low discriminating items. In this dissertation, this limitation of IPUI will be acknowledged and the properties of IPUI for 1PL will be investigated.

CHAPTER 4

RESEARCH QUESTIONS AND METHODS

4.1 Research Questions

As stated in the previous chapter, there is a need to evaluate item pools. This study focuses on creating a new index to evaluate the quality of an item pool for a given adaptive test and examinee population. This study investigates whether this index provides additional information about item pools on top of existing methods to evaluate a CAT (such as bias, SE of ability estimates, MSE, item pool information function). In addition, the methods to diagnose the item pools using this index are investigated. The research questions of this study are the following:

1. For a given population of examinees and adaptive test design, does this index changes as the item pool quality changes?
2. How does the magnitude of this index changes as the CAT specifications that affect the item pool quality changes?
3. How can this index be used to diagnose the shortcomings of the item pools and assist test developers to improve the quality of their item pools?
4. Can this index be used to evaluate the item pool quality of an operational CAT?
5. Can this index be used to diagnose the shortcomings of an operational item pool?

4.2 Research Methods

There are two phases of this study. In the first phase, the first three research questions are investigated using CAT simulations with different specifications. In the second phase of the

study, the last two research questions are examined for an existing operational CAT.

4.2.1 First Phase - Simulated Data

The first phase of the study consist of three sets of simulations to answer research questions 1, 2 and 3. In these three sets of simulations, generated data were used.

4.2.1.1 Common CAT Specifications

All adaptive tests in the first phase shared some common specifications. Simulations were performed using R programming language (R Core Team, 2014). The 1PL model with scaling parameter D = 1.7 was used. For the 1PL, the probability of correct response is:

$$P(u_{i_k} = 1) = \frac{1}{1 + e^{-1.7 \cdot \left(\theta - b_{i_k}\right)}}. \tag{4.1}$$

Each test started with an initial ability estimate of 0. Items were selected using MFI as explained in Section 2.3.2.1. For interim and final ability estimation, the EAP method with prior mean 0 and prior variance 4 was used at the beginning of the test until the examinee obtained at least one correct and one incorrect response. After a heterogeneous response string, ability was estimated using MLE. The variance of the prior distribution was chosen to be 4 (instead of a strong prior with variance 1 or even lower) to reduce the impact of it on the ability estimates and consequently the item selection.

In practice, ability estimates are usually confined within an arbitrary interval (Wang & Vispoel, 1998), for example between -4 and 4. This is usually done to deal with the infinite ability estimates of MLE for examinees with all correct or all incorrect response strings. Another reason for this practice is confining extreme ability values into a practical interval. In the first phase of the simulations, ability estimates were not confined into such an arbitrary interval. There were two reasons behind this decision. First, EAP estimation was used until an examinee obtained a heterogeneous response string. So, infinite ability estimates were not a problem. Second, since the first phase was a theoretical study, the aim was to observe the

effects of conditions on dependent variables without the interference of such arbitrary rules. For example, if there was an arbitrary limit for ability estimates, the standard deviation of the ability estimates at the extremes of the ability distribution might be depressed due to such rules. This would limit the generalizability of the conclusions from the analysis because it would not possible to strip out the effects of this arbitrary rule.

For all of the simulations in Research Question 1 and 2, 10,000 examinees were simulated for each condition. This number was enough to get a representative sample from the ability distributions and minimize the effects of sampling errors. Larger samples would not have much added value on top of this number due to the diminishing returns. And, since there were many conditions to simulate, computing time would increase exponentially. For Research Question 3, 1000 examinees at each θ value were simulated. Since there are 31 θ values between -3 and 3 with 0.2 interval, a total of 31,000 examinees were simulated for each condition.

To generate responses, first, a random number from a uniform distribution between 0 and 1 was generated. Then, the examinee's probability of correctly answering the item was calculated. If the random uniform number was smaller than the probability of correct response, a score of 1 was assigned as a response, otherwise a score of 0 was assigned.

4.2.1.2 Research Question 1

In the first set of simulations the utility of the IPUI was explored by checking whether this index changes systematically as the item pool quality changes, i.e. whether this index is sensitive to the changes in the quality of the item pool. The item pool quality was operationalized as:

1. The discrepancy between item pool distribution and examinee ability distribution
2. Item pool size

According to (1), item pool quality reduces as the discrepancy between the item pool distribution and ability distribution increases. According to (2), item pool quality decreases as the size of an item pool (i.e. the number of items in an item pool) decreases. It was hypothesized that the value of the index will decrease as the quality of the item pool decreases. Research Question 1 was answered by checking whether the IPUI changes as these two item pool quality indicators change.

Item pool and examinee ability discrepancy A CAT simulation was performed to check whether the IPUI decreases as the discrepancy between item pool distribution and examinee ability distribution increases. Even though there are many ways to increase the discrepancy between two distributions, in this study it was assumed that both item difficulty (b parameter) distributions of the item pools and the ability distribution of examinees were normally distributed with same standard deviations, 1. The discrepancy between the item pool and ability distribution was increased (or decreased) by increasing (or decreasing) the difference between the means of the item difficulty distribution and ability distribution.

In the simulation, the item difficulty distribution of the item pools were fixed to the standard normal distribution. On the other hand, ability distributions had different means ranging from -3 to 3 with 0.5 intervals. Standard deviations of the ability distributions were fixed to 1 as well. This setup allowed observation of whether the IPUI changes when the discrepancy between the item pool and examinee ability distribution takes values -3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5 and 3. It was hypothesized that the IPUI would take the highest value when the discrepancy between the two distributions was 0 and IPUI would decrease as the discrepancy moves either towards -3 or 3.

All of the CAT specifications were same for each of the 13 discrepancy conditions. The item pool was consist of 250 items. As mentioned above, the item pool had a standard normal distribution. The same item pool was used for all conditions. For each condition, 10000 examinees were simulated. CAT tests had a fixed test length of 20 items. Item pool size was

chosen according to this test length. As mentioned above, Stocking (1994) suggested that item pool size should be twelve times the length of adaptive test. There were no constraints on the item selection algorithm such as exposure control or content balancing.

Item pool For the second part of Research Question 1, whether the value of IPUI changes as the item pool size changes was explored. Item pool size is the second operationalized indicator for the item pool quality. It is hypothesized that as the item pool size increases, the IPUI index will increase as well. But instead of linear, this relationship is hypothesized to be a monotonic non-linear increase. To observe this relationship, CAT with different item pool sizes were simulated. There were 11 item pool size conditions: (1) very small item pool with the size of the test length, 20 in this case; (2-4) Small item pool sizes with 40, 60 and 80 items; (5-11) large item pools with sizes 100, 200, 300, 400, 500, 750 and 1000. Item difficulty parameters of all item pools were generated from standard normal distribution. For each condition, 10,000 examinees generated from standard normal distribution were used. Test length of all tests were 20. There were no constraints on the item selection algorithm.

When comparing item pools, especially for the ones with small numbers of items, the effect of sampling error is expected. For example, for the item pool of size 20, if 20 items from the standard normal distribution is randomly selected and the IPUI index is calculated based on this item pool, possibly this index will differ from another 20 items that are randomly chosen. To alleviate the effects of the sampling error, simulations were repeated 25 times for each item pool size condition and results were reported based on these replications. A larger effects of sampling error is expected for small item pool sizes. But for larger item pool sizes, the effect of the sampling error is expected to be disappear. Performing these replications was worthwhile because it showed the sensitivity of the index as well.

In this thesis, the 1PL model was used. So, the items differed only by their difficulty parameters. All items were assumed to discriminate examinees equally well, and examinees were assumed to not guess. On the other hand, in 2PL or 3PL models, items differ in their

discrimination parameters and their guessing parameters. In these models, a high quality item has a high discrimination parameter and low guessing parameter. High quality items provide more information about the examinee, given that the item difficulty is very close to the examinee's true ability level. In these models,increasing the size of an item pool without considering the quality of items might result in unexpected item pool performances. A small item pool with high quality items can perform better than a larger item pool with low quality items. Consequently, the size of an item pool will not be the only determinant of the quality of an item pool in 2PL or 3PL models.

For each condition, biases, standard errors, mean squared errors of the ability estimates and the fidelity coefficient (McBride, 1977), i.e. correlation between true and estimated abilities have been calculated. In addition, IPUI values and exposure rates of the items for each condition were calculated. For a CAT with test length K and N examinees:

$$\text{Mean Standard Error } = \overline{SE} = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{k=1}^{K_j} I_{i_{jk}}\left(\hat{\theta}_{jk}\right)\right]^{-1} \tag{4.2}$$

$$\text{Mean Squared Error} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right)^2 \tag{4.3}$$

$$\text{Mean Bias} = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j - \theta_j \tag{4.4}$$

$$r_{\theta\hat{\theta}} = \frac{\sum_{j=1}^{N}\left(\theta_j - \bar{\theta}\right)\left(\hat{\theta}_j - \bar{\hat{\theta}}\right)}{\sqrt{\sum_{j=1}^{N}\left(\theta_j - \bar{\theta}\right)^2}\sqrt{\sum_{j=1}^{N}\left(\hat{\theta}_j - \bar{\hat{\theta}}\right)^2}} \tag{4.5}$$

where θ_j is the true ability of examinee j, $\hat{\theta}_j$ is the estimated ability of examinee j, $\bar{\theta}$ is the mean of true abilities of N examinees, $\bar{\hat{\theta}}$ is the mean of estimated abilities of N examinees, $I_{i_{jk}}(\hat{\theta}_{jk})$ is the Fisher information of item i_{jk} at $\hat{\theta}_{jk}$ and $r_{\theta\hat{\theta}}$ is the fidelity coefficient (correlation between true and estimated abilities).

4.2.1.3 Research Question 2

There are many factors that affect the quality of the item pools besides the size of an item pool or the discrepancy between the item pool and examinee population distribution, such as CAT specifications. These factors might influence the quality of the item pools differently. Changing CAT specifications might have positive or negative impact on the utilization of the item pool. For a set of specifications, item pool quality might be sufficient but if these specifications changes, item pool quality might change as well even though the item pool itself does not change. Exposure control is a good example for this. An item pool might be sufficient for an adaptive test with no exposure control, but imposing an exposure control procedure might reduce the quality of the item pool. In Research Question 2, how different CAT specifications might influence the quality of the item pool was investigated.

The results for adaptive tests with different specifications were compared to see (1) how these changes affect the index, (2) whether this index captures the changes in the item pool quality caused by the changes in CAT specifications, (3) the relationship between this index and other outcomes of the adaptive tests (such as bias, SE, MSE etc.), (4) whether this index captures the item pool quality changes that other indicators of a CAT cannot capture. The last aim would show the added value of this index to the current literature.

In order to see the main effects of the test specifications on the IPUI and other CAT outcome indicators, CAT simulations in which only one specification changes at a time were performed. Two CAT specifications were the focus: test length and exposure control. Clearly, CAT specifications might also interact with each other. Changes in two or more specifications at the same time might have a different impact on the item pool quality compared to changing one specification at a time. The effects of the changes might not be additive and there might be interactions between specifications. In this study, only the main effects were investigated for simplicity reasons. Specifically, the effects of test length and exposure control on the quality of the item pool were investigated.

Test length First, the effects of test length on IPUI and other outcomes of the adaptive test were investigated. As indicated in the rule of thumb by Stocking (1994), there is a direct relationship between test length and item pool size. It is expected that, everything being equal, as test length increases, item pool quality decreases, and consequently the value of the IPUI decreases. This relationship was explored by simulating CATs with different test lengths. Eighteen different test length conditions were tested to see the effects of test length on item pool quality. These conditions were test lengths 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 100, 200, 300 and 400. The last condition basically administered every item in the item pool to a simulee. In this sense this condition was not very different than a linear test. The only difference was the ordering of the items. In a linear test the order of items are generally same for every examinee.

10,000 examinees were generated from a standard normal distribution. The same set of examinees were used for all conditions. The item pool size was 400, approximately 10 times the median of the test length conditions. The same item pool used for all conditions. Item difficulty parameters of the item pool were generated from standard normal distribution. Since the item pool size was rather large, it was assumed that the effect of sampling error was minimal. Consequently, there were no replications of the simulations with different item pools of the same size. There weren't any constraints on the item selection algorithm.

Exposure Control Exposure control is the second CAT specification imposed on the item selection that was explored. Exposure control is an important factor affecting the item pools especially in high stakes CATs. It is one of the main reasons that forces the test developers to increase the size of their item pools. For this reason, it is crucial to explore the interaction between IPUI and exposure control.

Exposure control has many variations in adaptive testing as explained in Revuelta and Ponsoda (1998), Georgiadou et al. (2007), Leroux, Lopez, Hembry, and Dodd (2013). This study focused only on randomesque exposure control procedure (Kingsbury & Zara, 1989)

due to its simplicity and widespread usage.

There were 12 item exposure conditions: (1) no exposure control, (2-11) randomesque with 3, 5, 7, 10, 13, 15, 20, 25, 50, 100 items and (12) total-random-selection of items. For example, in randomesque with 10 items, one item out of the 10 most informative items at examinee's current ability estimate was selected. In total-random-selection of items, items were randomly selected out of all available items, regardless of their information value. This case served as the worst case scenario in the sense of the efficiency of a CAT among the exposure control conditions. It was expected that, as more randomness was imposed on the item selection procedure, the value of the IPUI would decrease.

As in the previous simulations, everything except the exposure control were the same in the CAT simulations between the conditions. For each condition, 10,000 examinees were generated from a normal distribution with a mean of 0 and a standard deviation of 1. Test length was 20 for all conditions. Item pool consisted of 250 items where item difficulties were generated from a normal distribution with a mean of 0 and a standard deviation of 0.7. IPUI values were compared at each condition along with mean SE, MSE, mean bias, fidelity coefficient and exposure rates.

4.2.1.4 Research Question 3

Research Question 3 focuses on the usefulness of this index as a diagnostic tool for item pools. The scenario of this research question was hypothesized as the following.

Suppose a state testing agency is planning to develop a test to measure the end of year achievement levels of students. The purpose of the test is to measure each student as precisely as possible. Students had a wide range of abilities. CAT is decided to be the best way to achieve this goal. The state testing agency wants to measure the students in three content areas. Content area 1 (i.e. arithmetic) is generally regarded as easy among the average students and there are not too many difficult items in this content area. Content area 2 (i.e. algebra) is regarded as medium difficulty and content area 3 (i.e. trigonometry) is regarded

as difficult among the average students. The state testing agency has three CAT plans to compare before implementing the test. Previous item development efforts showed that the item difficulty parameters of the items from content area 1 had a normal distribution with mean difficulty -1 and standard deviation 0.3. The item difficulties of items from content area 2 had a normal distribution with mean difficulty 0 and standard deviation 0.3. The item difficulties of items from content area 3 had a normal distribution with mean difficulty 1 and standard deviation 0.3. In the following simulation plans, item difficulty parameters of the item pools were generated from these distributions for each content area.

Here are the three test plans investigated for this hypothetical state testing agency to demonstrate the use of IPUI:

Plan 1. For this plan, an item pool of 90 consisting of an equal number of items from each content area (30 items each) was created. The CAT specifications of this item pool were the same as the common specifications of the previous research questions (see Section 4.2.1.1 on page 37). Length of the test was 15 items. Different than the previous CAT specifications, in this test plan, content balancing was imposed on the item selection algorithm. Examinees should respond to exactly 5 items from each of the three content areas. After the administration of each item, the content area that has the largest discrepancy with the target value (i.e. 5) was selected. The most informative item within this content area at the examinee's intermediate ability level was administered.

Plan 2. In this plan, the same item pool created for the first plan was used. The test specifications of this plan were the same as the first plan except for the content balancing. For the second plan, no content balancing was imposed on the item selection algorithm. The most informative item at the examinee's intermediate ability estimate was administered regardless of the content of the item.

Plan 3. The third plan involves the creation of a CAT test for each content area separately

with a larger item pool. Only the results of the third content area were compared to the other plans in lieu of the other content areas. An item pool of 90 items from the distribution mentioned above for content area 3 was generated. Since there is only one content area, there was no content balancing for this test. The CAT specifications of this plan were the same as Plan-2.

The size of the item pools for each plan was the same, 90 items. The item pools of Plan 1 and 2 were the same, but they differed by the content balancing imposed on the item selection algorithm. The CAT specifications of Plan 2 and 3 were the same but they differed by the distribution of the item difficulties of the item pools. The results of Research Question 3 shows the use of IPUI as a diagnostic tool. Using this diagnostic information, this hypothetical state testing agency can decide which plan provides the best measurement option given the practical constraints. They can also see the weak points of the item pools and choose a plan accordingly, or decide to improve the item pools.

To compare each plan, 1000 examinees were simulated at each θ point between -3 and 3 with 0.2 intervals. The number of items and the distributions of the item pools are shown in Table 4.1.

Table 4.1: Item Pool Information for Research Question 3

	Content Area 1	Content Area 2	Content Area 3	Total
Plan 1-2 (IP 1-2)	30 ∼N(-1, 0.3)	30 ∼N(0, 0.3)	30 ∼N(1, 0.3)	90
Plan 3 (IP-3)	0	0	90 ∼N(1, 0.3)	90

Note. The number of items and the distribution of items are given within each cell. Numbers within the parentheses are the means and the standard deviations of the generating distributions.

The mean values of bias, SE, MSE and IPUI were calculated at each true θ condition for each plan. The diagnostic utility of IPUI was investigated and compared with the diagnostic utilities of other CAT outcomes.

4.2.2 Second Phase - Real Data

In the second phase, the simulations were based on the real item pools of an operational CAT. National Council of State Boards of Nursing (NCSBN) provided the operational item pools of National Council Licensure Examination for Registered Nurses (NCLEX-RN) exam. NCLEX-RN (National Council of State Boards of Nursing [NCSBN], 2012) is a nursing licensure exam administered by NCSBN. NCLEX-RN examination "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, 2012, p. 3). Exam is delivered via CAT.

Description of NCLEX-RN Exam The specifications of the CAT algorithm used by NCLEX-RN is complex. Items that pass the quality control are calibrated using the Rasch model. Item pools are used for a certain period of time and renewed with a new item pool due to security reasons. Item pools consist of 8 content areas. Examinees should answer a specified proportion of questions from each content area. The content distribution of the examination is in Table 4.2 (NCSBN, 2012, p. 5).

Table 4.2: Distribution of Content for NCLEX-RN Examination

Content Area	Percentage of Items
Safe and Effective Care Environment	
• Management of Care	17-23%
• Safety and Infection Control	9-15%
Health Promotion and Maintenance	6-12%
Psychosocial Integrity	6-12%
Physiological Integrity	
• Basic Care and Comfort	6-12%
• Pharmacological and Parenteral Therapies	12-18%
• Reduction of Risk Potential	9-15%
• Physiological Adaptation	11-17%

The CAT procedure starts with an initial ability estimate that is lower than the average ability estimate of the examinees. Initially, Owen's Bayesian procedure (Owen, 1975) is used

for ability estimation and item selection until the examinee has at least one correct and one incorrect response in her response string. A normal prior distribution with mean 0 and variance 4 is used. After the examinee has at least one correct and one incorrect response in her response string, ability is estimated using MLE. The maximum value of the likelihood function is found using the Newton-Raphson algorithm.

For each question, the content area of the item is selected first. The content area which deviates most from the target content proportions is selected. Among the available and unadministered items within the selected content area, one of m items which have the maximum amount of information at the current ability estimate is randomly selected and administered to the examinee. Here, m is the parameter for the randomesque exposure control. After the examinee responds the item, her ability and the standard error of her ability is updated.

After the examinee completed 60 items, the CAT program checks whether the cut score is contained within a 90% confidence interval around the ability estimate. If the cut score is outside the confidence interval, the test is terminated. If the ability estimate is above the cut score, the examinee passes the test, otherwise fails.

If the cut score is within the 90% confidence interval around the ability estimate, the test continues until the cut score falls out of the confidence interval. The CAT program continues to administer items up to the maximum test length of 250, if the cut score is still within the examinees estimated ability confidence interval. After the administration of 250th item, the test is terminated. The examinee passes the test if the final ability estimate is greater than the cut score, otherwise the examinee fails.

4.2.2.1 Research Question 4

NCSBN administers NCLEX-RN throughout the year. Due to security reasons, test developers change the item pools within certain intervals. Each item pool is selected to meet the test specifications and passes through quality control. The index proposed in this thesis can help

the test developers to check the sufficiency of the selected item pool.

For this thesis, NCSBN provided the item parameter values, content area information of the retired item pools, the ability distribution of the previous test takers and the test specifications required to mimic the CAT procedure. In addition, NCSBN provided response strings of five anonymous test takers. This information was used to ensure that the CAT procedure written for this dissertation indeed mimics the real CAT algorithm used operationally.

For this research question, the quality of five operational item pools were investigated. On top of these five operational item pools, four additional item pools were generated. Two of them were ideal item pools developed for the specifications and target population of the NCLEX-RN test to form a baseline for comparisons. One of these ideal item pools created using a fixed bin size of 0.4, the other is created using a fixed bin size of 0.8. These bin sizes are the same bin sizes used in He and Reckase (2013). The details of the development of the ideal item pools is described in Section 4.2.2.1 on the following page.

The third item pool created is called the half-item-pool (Half-IP). This item pool has half the size of the first operational item pool. For each content area, half of the items were randomly selected and then removed from the first operational item pool. The remaining items formed the half-item-pool. A fourth generated item pool is called one-third-item-pool (One-Third-IP). This item pool was created in a similar way to the half-item-pool. Instead of half of the items, two thirds of the items randomly removed from each content area. The remaining items formed the one-third-item-pool. These item pools are the other extremes of the ideal item pools.

Evaluation of these item pools were performed using CAT simulations for each of these eight item pools. Normally distributed 50,000 examinees were simulated. The mean and the standard deviation of the normal distribution was the same as the ability distribution of the real examinees. The same 50,000 examinees were used for each item pool condition. IPUI was calculated for each item pool condition. Additionally, for each item pool, mean standard error (Equation (4.2)), mean squared error (Equation (4.3)), mean bias (Equation (4.4)),

fidelity coefficient (Equation (4.5)) and the percentage of correct classification was calculated and compared to the IPUI. Besides these outcome variables, since NCLEX-RN test is a variable length test, the average test length and the mean exposure rates were also calculated.

Ideal item pool creation Ideal item pools were created using the bin-and-union method as described at Reckase (2010) and He and Reckase (2013). According to van der Linden et al. (2006) an ideal item pool should

> . . . consist of a maximal number of combinations of items that (a) meet all content specifications for the test and (b) are most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for a population of examinees. (p. 82)

The bin-and-union method uses this principle to build an ideal item pool for a given examinee population and a set of test specifications. For this dissertation, two ideal item pools for an examinee group with the same ability distribution as the real examinee group were created. The CAT specifications were the same as the NCLEX-RN examination. Initially, 10,000 examinees were simulated from a distribution that represents the real examinee population. For these examinees, a CAT test was performed with the assumption of maximally informative item is available at each stage of the adaptive test, whatever the value of intermediate ability estimate is. Since the Rasch model was used in this study, b values of the maximally informative items were equal to the intermediate ability estimates.

Initially, bins within the ability scale with the following range limits were created:

Ideal item pool with bin size 0.4: $(-\infty$, -3), [-3, -2.6), [-2.6, -2.2), [-2.2, -1.8), [-1.8, -1.4), [-1.4, -1), [-1, -0.6), [-0.6, -0.2), [-0.2, 0.2), [0.2, 0.6), [0.6, 1), [1, 1.4), [1.4, 1.8), [1.8, 2.2), [2.2, 2.6), [2.6, 3), [3, $\infty)$

Ideal item pool with bin size 0.8: $(-\infty$, -2.8), [-2.8, -2), [-2, -1.2), [-1.2, -0.4), [-0.4, 0.4), [0.4, 1.2), [1.2, 2), [2, 2.8), [2.8, $\infty)$

These bins were used to tally the number of items required for each range. The number of items within each bin was set to 0 at the beginning of the simulation. Then, for each examinee, a CAT test was simulated and the b values of the items that were administered were recorded. At the end of the test, for each content area, items were distributed to bins according to their b values. For each bin range, if the number of items in that bin range was larger than the previous number of items assigned to that bin, then this larger number was assigned to that bin. At the end of the simulations for the entire examinee group, the maximum number of items that were necessary for each bin range was obtained. This gave the distribution of the ideal item pool.

After obtaining the distribution of the ideal item pool, actual ideal item pools were generated. A random number from the standard normal distribution was generated. This random number was assigned to the appropriate bin if that bin did not reach its maximum size yet. This procedure continued until each bin was filled. This way, an ideal item pool for this examinee group was obtained.

4.2.2.2 Research Question 5

In this research question, the utility of the IPUI as a diagnostic tool for an operational CAT was investigated. Analyses in this research question were the extension of the Research Question 3 to an operational CAT with different specifications.

In this research question, 6 item pools were studied. Two of these item pools were the operational item pools, two of them were the ideal item pools and the last two item pools were One-Third and Half item pools created for the previous research question. Even though there were five operational item pools available, only two of them were investigated. The results of the Research Question 4 showed that the differences between the operational item pools were minimal. For this reason, only the first two operational item pools were investigated further.

For each item pool condition, 1000 examinees at true θ values -3, -2.8, -2.6, -2.4, -2.2, -2, -1.8, -1.6, -1.4, -1.2, -1, -0.8, -0.75, -0.7, -0.65, -0.6, -0.55, -0.5, -0.45, -0.4, -0.35, -0.3, -0.25,

-0.2, -0.15, -0.1, -0.05, 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8 and 3 were simulated using the NCLEX-RN CAT specifications. Outside the θ interval $(-0.8, 0.8)$, the difference between the simulated θ values was 0.2. Between $[-0.8, 0.8]$, the difference between the simulated θ values was 0.05 because these points were close to the cut score and it was important to make more precise diagnosis close to the cut score where the decision accuracy was paramount. At each θ value, the mean of IPUI values, mean SE, MSE, mean bias and the decision accuracy was calculated. Each outcome variable was compared to the IPUI to show the diagnostic utility of IPUI. The distributions of IPUI values at each conditional θ value gave a local diagnosis of the item pool.

CHAPTER 5

RESULTS

The results of the analyses are divided into two phases. In the first phase, the results of the first three research questions are presented. In the second phase the results of the fourth and fifth research questions are presented.

5.1 First Phase - Simulated Data

In this first phase of the results, results of first three questions are presented. These three questions are based on simulated data. Test specifications were common in these three research questions except the conditions being tested (Section 4.2.1.1). The design of the first phase was kept as simple as possible to demonstrate the effect of the IPUI. A more complicated CAT design was investigated in the second phase of the study.

5.1.1 Research Question 1

In Research Question 1, item pool quality is operationalized as (1) the discrepancy between examinee ability distribution and item difficulty distribution of the item pool, and (2) the item pool size. In the following two sections the results of the simulations examining these two item pool quality indicators are presented.

5.1.1.1 Item Pool and Examinee Ability Discrepancy

There were thirteen conditions for item pool and examinee ability discrepancy. Both item difficulty parameters in the item pool and examinee abilities were generated from normal distribution with standard deviation of 1. The item pool was fixed for all conditions and the item difficulties had a standard normal distribution. The item difficulty distribution of

item pool is in Figure A.1 on page 167. The conditions differed by their means which ranged from -3 to 3 with 0.5 intervals. 10,000 examinees were simulated for each condition. The distributions of true thetas are plotted in Figure A.2 on page 168. For each condition, mean of bias, SE, MSE and IPUI were calculated.

The results of the simulations are summarized in Figure 5.1 and Table 5.1. Figure 5.1 shows the change in bias, SE, MSE and IPUI for each discrepancy condition. Table 5.1 shows the means and standard deviations of these output variables.

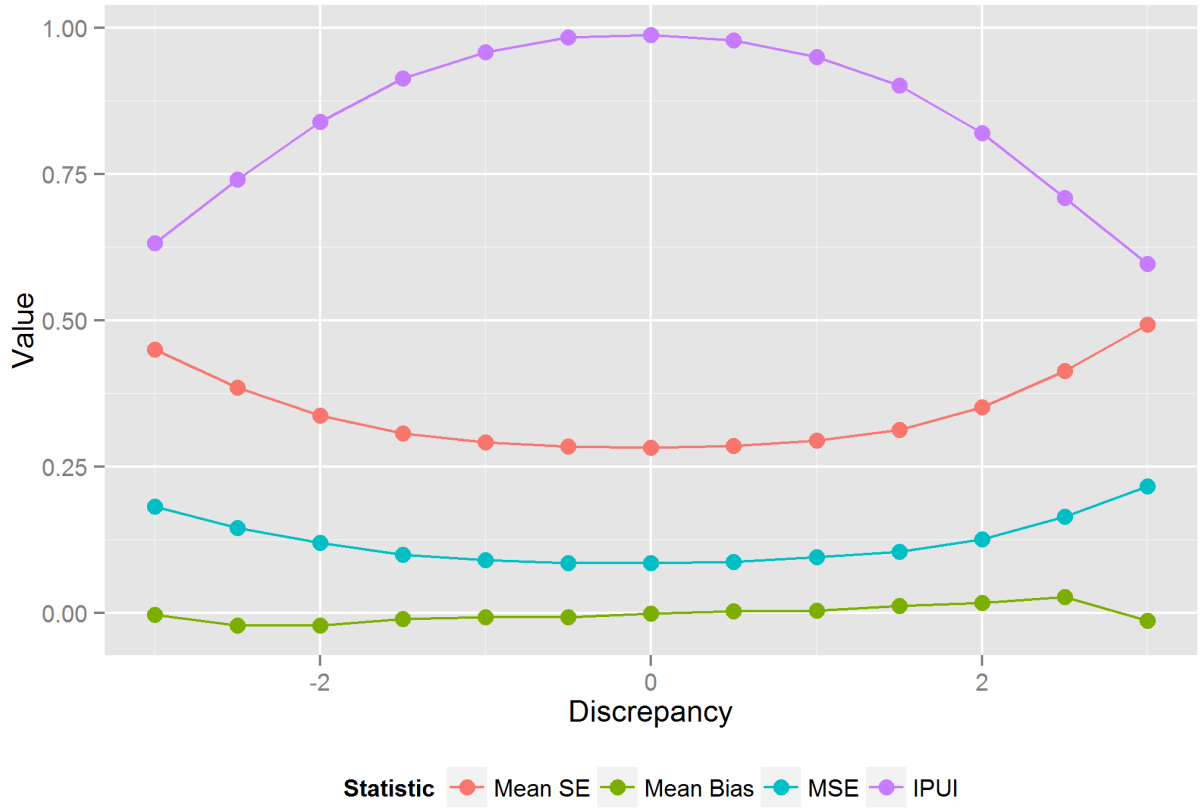

Figure 5.1: Summary Statistics for Research Question 1 - Discrepancy between Item Difficulty Distribution of Item Pool and θ Distribution

Bias Figure 5.1 does not show a clear relationship between the bias of ability estimates and the amount of discrepancy between ability distribution and item pool. Means of the biases were close to 0 for all conditions. On the other hand, the standard deviations of biases given

Table 5.1: Summary Statistics for Research Question 1 - Discrepancy between Item Pool and Ability Distribution

Discrepancy	SE	Bias	MSE	IPUI
-3.0	0.451 (0.190)	-0.002 (0.426)	0.182 (0.312)	0.632 (0.264)
-2.5	0.385 (0.157)	-0.021 (0.381)	0.145 (0.265)	0.741 (0.255)
-2.0	0.337 (0.116)	-0.020 (0.346)	0.120 (0.212)	0.839 (0.219)
-1.5	0.307 (0.075)	-0.009 (0.317)	0.100 (0.175)	0.913 (0.166)
-1.0	0.292 (0.046)	-0.007 (0.301)	0.091 (0.143)	0.958 (0.114)
-0.5	0.284 (0.026)	-0.007 (0.292)	0.085 (0.124)	0.983 (0.068)
0.0	0.283 (0.024)	-0.001 (0.293)	0.086 (0.131)	0.987 (0.060)
0.5	0.285 (0.035)	0.003 (0.296)	0.087 (0.142)	0.979 (0.084)
1.0	0.295 (0.065)	0.004 (0.309)	0.096 (0.163)	0.950 (0.135)
1.5	0.314 (0.096)	0.013 (0.323)	0.105 (0.185)	0.901 (0.186)
2.0	0.352 (0.144)	0.017 (0.356)	0.127 (0.229)	0.820 (0.239)
2.5	0.414 (0.189)	0.028 (0.405)	0.165 (0.297)	0.710 (0.272)
3.0	0.493 (0.220)	-0.013 (0.466)	0.217 (0.375)	0.596 (0.274)

Note. Numbers within the parentheses are standard deviations of each outcome. SE: Standard Error; MSE: Mean Squared Error

in Table 5.1 displays a pattern. As the discrepancy between ability distribution and item pool increased, the standard deviation of the biases increased as well. For low discrepancy conditions the variability of biases were low. Most of the examinees had biases close to 0. For high discrepancy conditions the variability of the biases were higher. Figure A.3 on page 169 also shows this visually.

Standard Error Figure 5.1 shows that as the absolute value of discrepancy between the item pool and ability distribution increased, SE of the ability estimates increased as well. When there was a close match between ability distribution and item pool, the overall SEs of examinees were low. For large discrepancy conditions, the mean of the SEs were larger.

Similar type of increase can be observed for the standard deviations of the SE values. Table 5.1 shows that as the absolute value of discrepancy increased, the variability of standard errors increased. The variability of SEs are also plotted in Figure 5.2. In Figure 5.2, each point represents the SE value of a simulated examinee. Each examinee is plotted close to

the discrepancy condition it belongs. The points are scattered around the exact discrepancy value to show the distribution of the values. In addition to points, a box-plot was added to show the shape of the distribution. Only the box and whiskers of the box-plots were plotted because the outliers are already shown by the points in the graph. As box-plots and the distribution of the points show, SE for each condition had a positive skew.

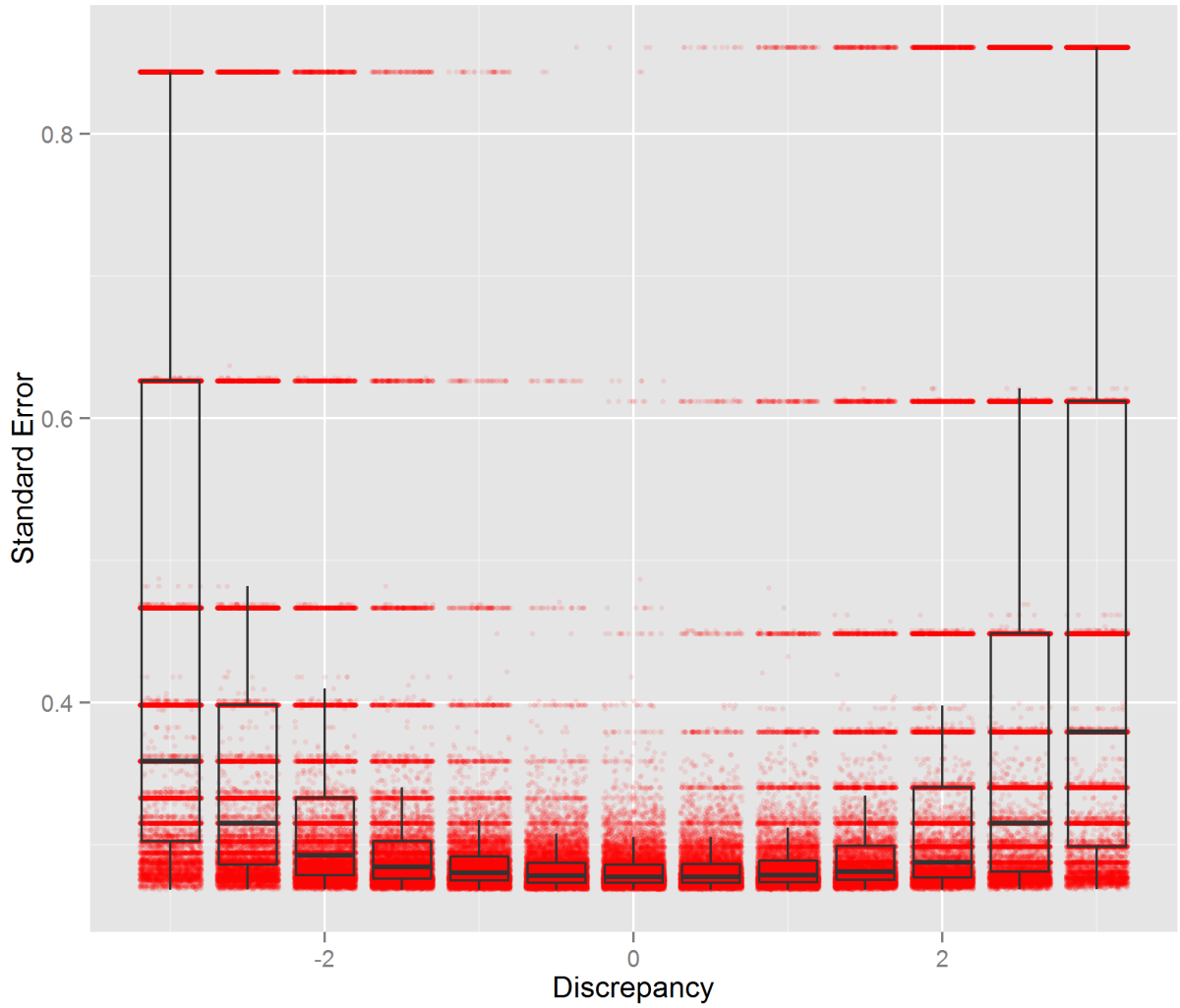

Figure 5.2: Distribution of Standard Errors within each Discrepancy Condition

Mean Squared Error The relationship between MSE and discrepancy conditions were similar to SE. As discrepancy between item pool and ability distribution increased, mean of

the MSEs increased as well. For small discrepancy conditions the MSE distributions were almost identical. Table 5.1 shows that both means and standard deviations of MSE values were very close when the discrepancies were below 1 unit. Figure A.5 on page 171 shows the distribution of MSE for each discrepancy condition. As can be seen from this plot, the variation in MSE values increased as the discrepancy increased. The distributions of MSE were positively skewed as SE distributions.

IPUI It can be clearly seen in Figure 5.1 that as the absolute value of discrepancy between the item pool and ability distribution increased, the IPUI values decreased. When there was no discrepancy or it was minimal, the mean of IPUI was almost 1. This means, for almost all of the examinees item pool was able to provide appropriate items throughout the test. Item pool was sufficient for these examinee groups. But for high discrepancy conditions IPUI values decreased markedly. The mean of IPUI values went down to 0.6 for these examinee groups. For these examinees, item pool failed to provide appropriate items.

In addition to the decrease in IPUI values for high discrepancy conditions, the variability of the IPUI values increased for high discrepancy conditions. When there was a close match between item pool and examinee distribution, the standard deviation of the IPUI was small, around 0.06. The standard deviations of the IPUI values go up to 0.274 for high discrepancy conditions. This variability can also be observed from Figure 5.3. For all conditions, IPUI values range from 0.274 to 1. The variability increased as the discrepancy increased and distributions showed a negative skew for almost all conditions. But it can be observed that the skewness of the distributions went from negative to positive as the discrepancy increased.

Figure 5.4 shows the relationship between SE and IPUI for each discrepancy condition. Linear regression line was fitted to each plot. The correlation between SE and IPUI was printed at the bottom of each plot. For each discrepancy condition there was a negative relationship between SE and IPUI. But this relationship was not linear. Instead it was curvilinear. This graph clearly shows that SE and IPUI are not the functions of each other.

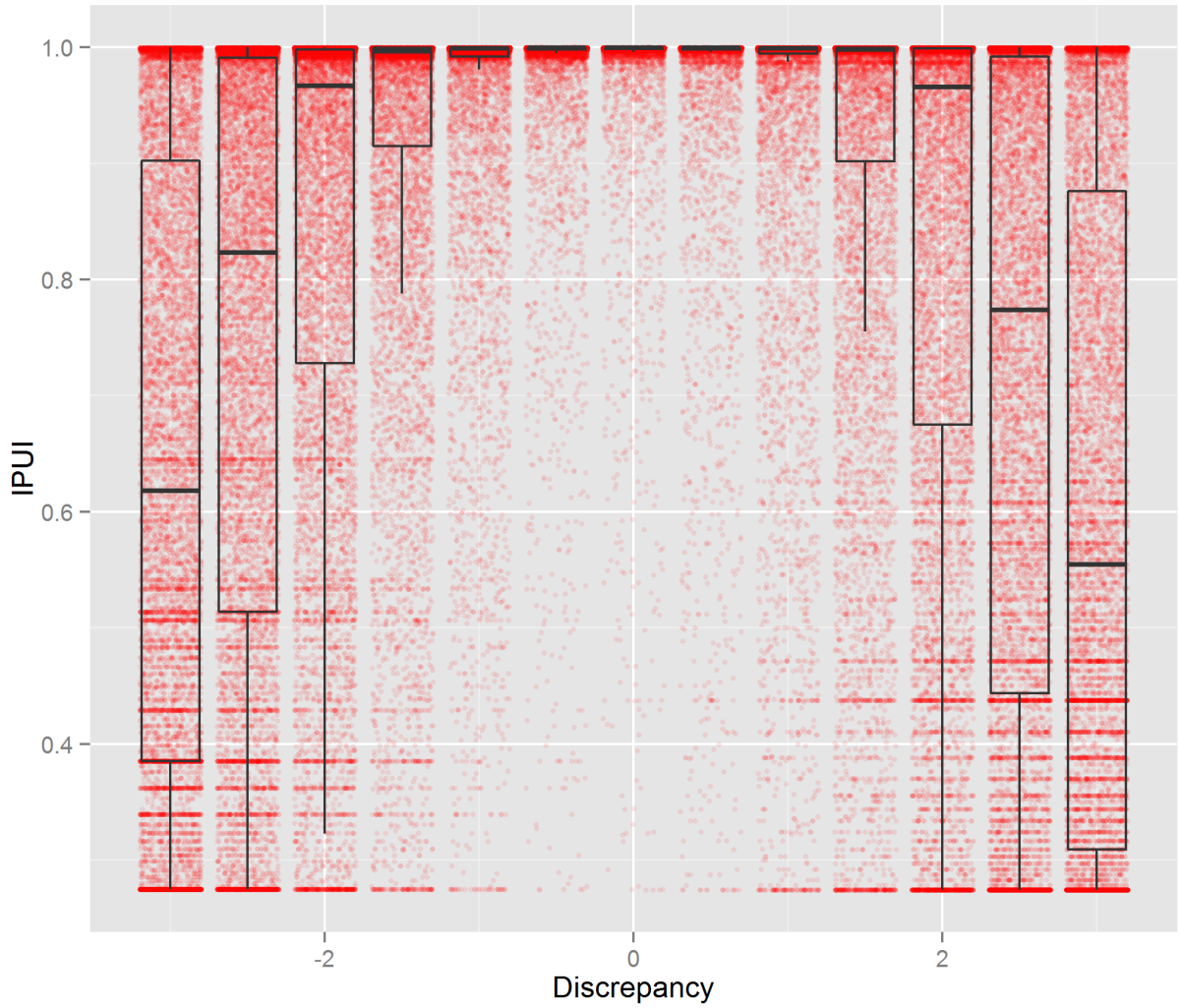

Figure 5.3: Distribution of IPUI within each Discrepancy Condition

Figure 5.4 shows that there were many cases where both IPUI and SE were low. This means the item pool did not provide optimum items to examinee but still the error of ability estimate is low. On the other hand, there were very few cases where both IPUI and SE were high. This means, when the item pool able to present optimum items to the examinee, the standard errors of the ability estimates will not be high.

Figure A.4 on page 170 shows the relationship between IPUI and bias for each discrepancy condition. There was a positive relationship between IPUI and bias when the discrepancy between item pool and ability distribution was negative. For positive discrepancy conditions

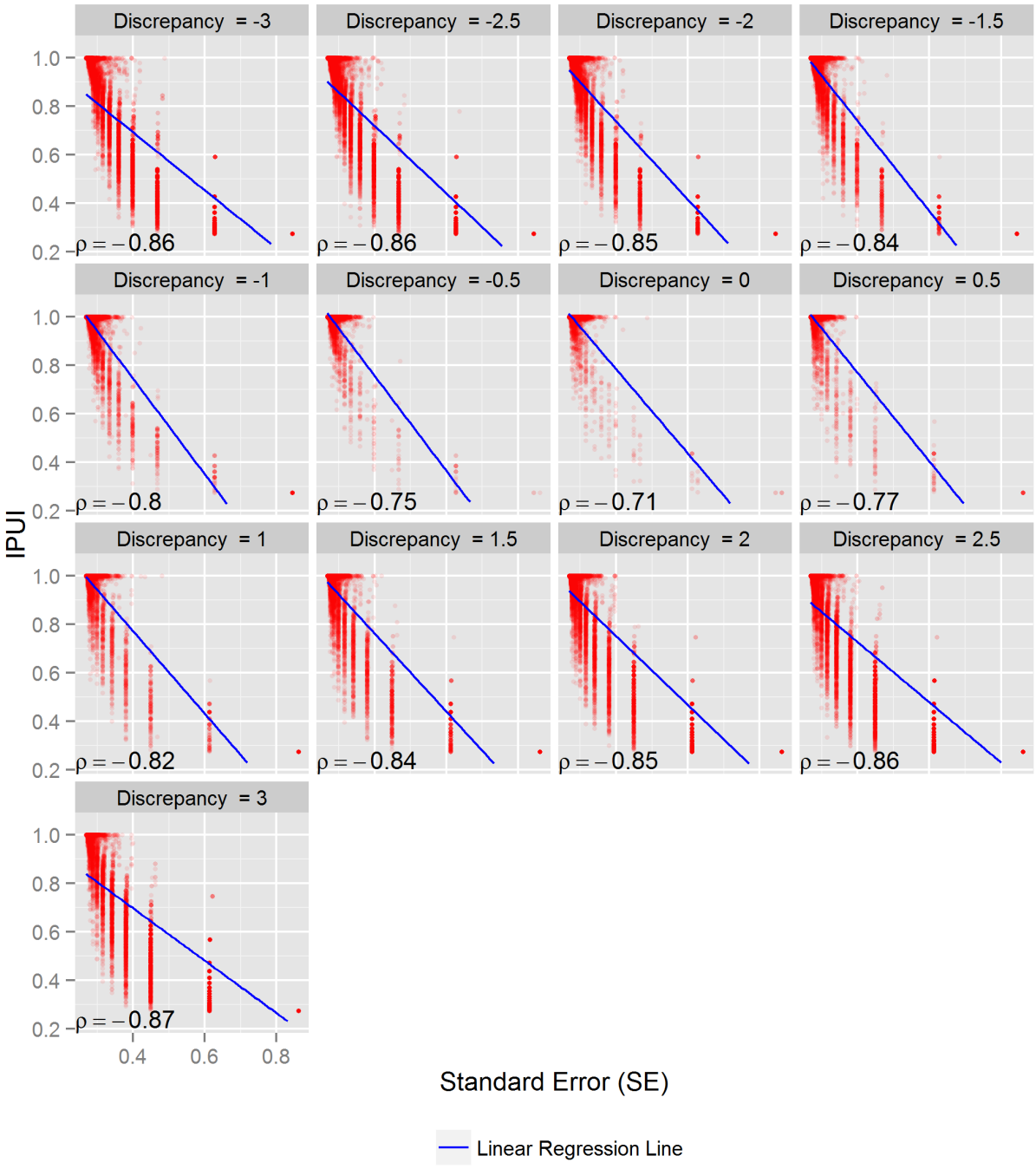

Figure 5.4: Relationship between Standard Error and IPUI for each Discrepancy Condition

IPUI and bias had a negative relationship. These relationships were not strong as the correlation coefficients at the bottom of each plot shows. This relationship can be explained by the regression to the mean effect.

Figure A.6 on page 172 shows the relationship between IPUI and MSE for each discrepancy condition. There was not an apparent pattern between these two outcomes.

One interesting observation in Figure 5.2 is the clustering of SE values. For example, when the discrepancy condition was -3, the three highest SE values were 0.84, 0.63 and 0.47. There were many examinees with these SE values. Out of 10,000 simulated examinees, the number of examinees with these SE values were 1362, 415 and 211 respectively. These SE values belong to the examinees with 0, 1 and 2 total correct responses respectively, who answered the same items in the same order. For 1362 examinees without any correct response, the CAT procedure administered the same items because of their responses. All of these items were the easiest items of the item pool. Consequently, their final ability estimates were the same as well. This is why their standard errors were the same. Even though there were more than 415 examinees with total correct score of 1, these 415 examinees made the same one error to the same question which result in the administration of same test items to these examinees. Same thing was true for 211 examinees with the third highest SE.

Similar phenomenon can be observed to some extent for IPUI values in Figure 5.3. The IPUI values had more variability in the values it took compared to SE. For discrepancy condition -3, there were 1383 examinees who had the same IPUI value. This number is a little larger than the number of examinees with same SE (1362). These 21 had different SE because their last item is correct. That last response did not change the IPUI value but changed the SE value for these examinees. Rest of the 1362 examinees did not have any correct response.

For the examinees who had two correct responses and the same SE values, the IPUI values were not always the same. If the location of these two errors were different, then the IPUI values would be different. The CAT processes of two such examinees is in Figure A.7 on page 173. Even though these two examinees had same SE, their IPUI values were different because of the position of their mistakes. This phenomenon is the main reason why there were so many different IPUI values compared to the SE. For examinees with only two correct

responses, SEs will be the same if the same items are delivered regardless of the location of the correct responses. Because, for Rasch model, the number of total correct responses is a sufficient statistic (Leeuw & Verhelst, 1986). If the CAT algorithm administers same items to two examinees but examinees correctly answer two different items, then their final score will be the same. Consequently their SEs will be the same as well. But for IPUI, if the location of the correct responses change, then the IPUI values will change as well. Because IPUI is the function of all intermediate ability estimates and the item parameters. On the other hand, SE is the function of final ability estimate and the item parameters.

For other discrepancy conditions similar thing happened. For discrepancy condition 3, many examinees got all items correct. As a result, the CAT algorithm administered the same items to these examinees and their final SEs came out the same. In Figure 5.2, it can be seen that the SE values for all correct and all incorrect answers were different. One important note is that, the item pools used in all conditions were same. So, if an examinee correctly answers all items in two different discrepancy conditions, then their standard errors would come out the same. By coincidence, the IPUI values of all incorrect and all correct items were very close for this item pool (0.2749047 vs. 0.2742634 respectively). As a result, the minimum values of IPUI values in Figure 5.3 for discrepancy condition -3 and 3 looked the same.

5.1.1.2 Item Pool Size

In the second part of the Research Question 1, item pool quality was operationalized as the size of the item pool. It was hypothesized that as the item pool size increases, the quality of the item pool increases as well. Here, the item pool parameter distributions were assumed to remain the same. Consequently, the values of IPUI were hypothesized to increase as the item pool sizes increase. This hypothesis was tested using 11 item pool size conditions. Item pools with 20, 40, 60, 80, 100, 200, 300, 400, 500, 750 and 1000 items generated. The item difficulty parameters (b-parameter) of all item pools were generated from a standard normal distribution. For each item pool size condition, there were 25 replications to observe

the effects of sampling error especially for the small item pools. For very small item pools, even though the item difficulties were drawn from a standard normal distribution, the items generated in one replication might have different item difficulties compared to the items generated in another replication. The item difficulty distributions of item pools for 19th replication is in Figure B.2 on page 175.

10,000 examinees were simulated from a standard normal distribution. The same set of examinees used for each replication and condition. The distribution of true θ is shown in Figure B.1 on page 174. Test length for the adaptive tests was 20 and there were no constraints on the item selection algorithm. The rest of the CAT specifications were the same as the common CAT specifications mentioned in Section 4.2.1.1 on page 37. For each condition and replication, bias, SE, MSE, exposure rates and IPUI values calculated.

Bias Figure 5.5 shows the mean bias distribution for each condition. Each point represents the mean of the biases of 10,000 simulees for each replication within a condition. Except the very small item pools (20 and 40), mean biases were very small for each condition. The values for replications were spread around 0 and does not show a pattern for item pool sizes larger than 40. For very small item pool sizes, the spread of mean bias values was higher. Figure B.3 on page 176 shows the bias distribution for the 19th replication of each condition. The spread of biases were also higher for very small item pools. For the rest of the item pool size conditions, the distributions of the biases were almost the same, normally distributed with the same means and standard deviations.

Standard Error Figure 5.6 shows the mean SE values for each replication within a condition. As the item pool size increased the mean SE decreased. For item pools that are larger than 200, the mean SE values were almost same. The spread of mean SEs within condition were higher for small item pools. This reflects the effect of sampling error. Figure B.4 on page 177 shows the SE distribution for the 19th replication of each condition.

Figure 5.5: Mean Bias Distribution by Item Pool Size Condition

The mean of SE for each condition decreased as the item pool size increased. After the item pool size of 100, the difference between means were very small. The spread of SE decreased as the size of the item pool increased. After 200 items the spread was almost the same for all conditions. But the number of outliers decreased as the item pool size increased. For item pool of size 1000, there was only one simulee with SE larger than .4. For other conditions this number increased as the item pool size decreased. Also the distributions were positively skewed for each condition.

Mean Squared Error Figure 5.7 shows the mean MSE values for each replication within a condition. MSE distribution shows a similar pattern as SE, as the item pool size increased the mean of MSEs decreased. The spread of the mean MSE values within a condition were larger

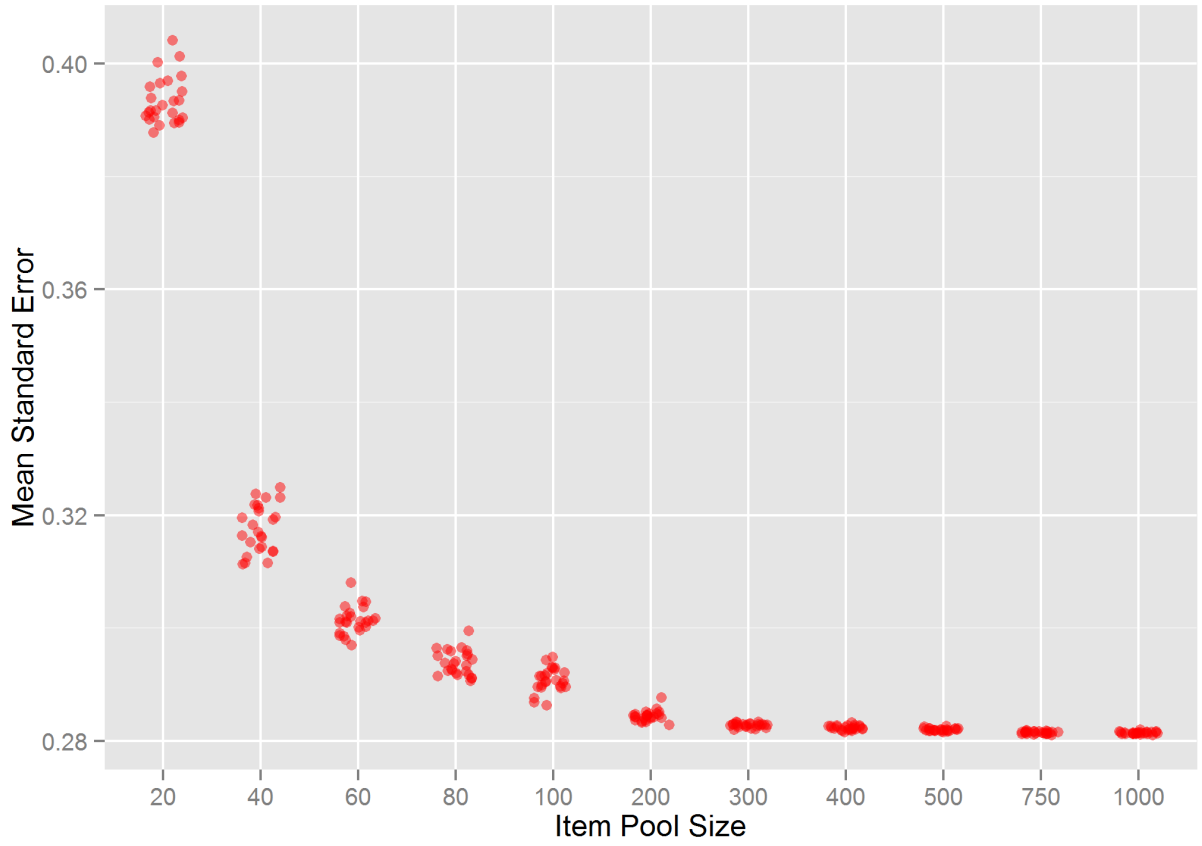

Figure 5.6: Mean Standard Error Distribution by Item Pool Size Condition

for small item pool size conditions and decreased as the item pool size increased. For item pools larger than 200 items, the spread was almost the same. Figure B.5 on page 178 shows the MSE distribution for the 19th replication of each condition. Even though the means and the spread of MSE values were larger for very small item pool sizes, the distributions were very similar for item pools that are larger than 40. All of the distributions were positively skewed.

Exposure Rates Mean exposure rate for replications were not calculated because the mean exposure rate of an adaptive test equals to the ratio of test length to the item pool size. Instead, Figure 5.8 shows the exposure rates of the items for the 19th replication of each condition. Since test length was fixed, the exposure rates of small item pools were overall

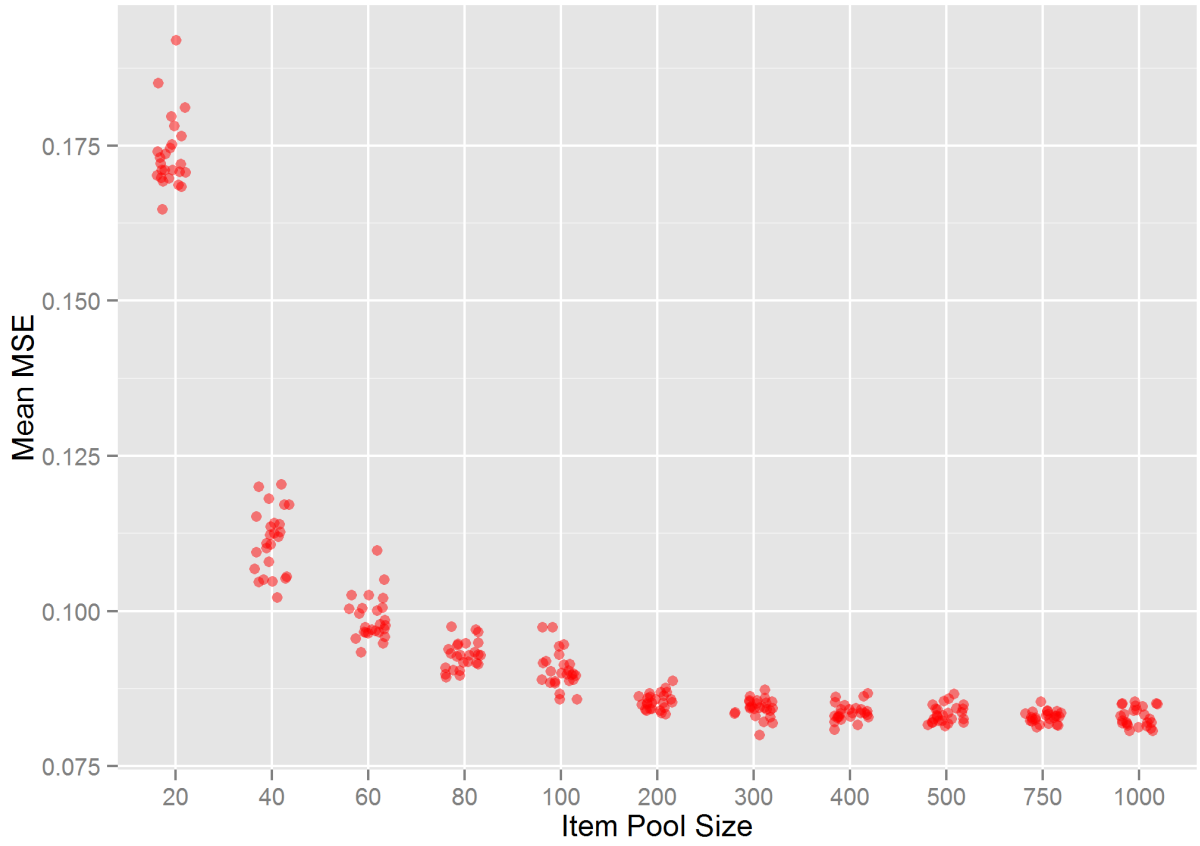

Figure 5.7: Mean Squared Error Distribution by Item Pool Size Condition

higher than the exposure rates of the larger item pools. Also, initial ability was same for all examinees. This led the selection of the same one item for each adaptive test. For this item, the exposure rate was 1. Depending on the correct or incorrect response, examinees were routed to the same second items. Exposure rates for these items were around .5 for large item pools and even higher for smaller item pools. Besides such items, for item pools that were larger than 200 items, majority of the exposure rates were lower than 0.2, the recommended maximum value for item exposure.

IPUI Figure 5.9 shows the mean IPUI values for each replication within a condition. There was a clear relationship between item pool size and mean IPUI values. As the item pool size increased, the mean values of IPUI increased as well. The spread of mean IPUI values were

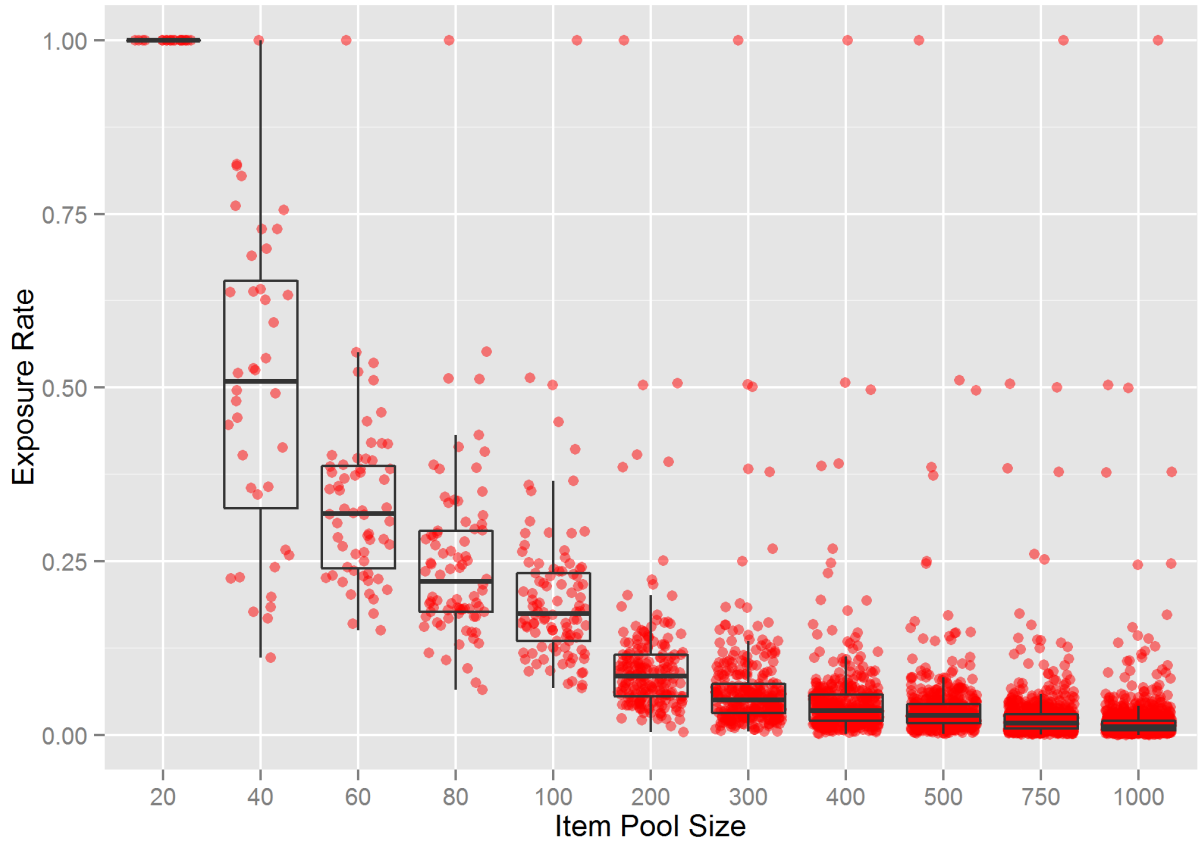

Figure 5.8: Exposure Rates by Item Pool Size Condition for Replication 19

larger for smaller item pool sizes, but the spread decreased as the item pool size increased.

Table 5.2 shows the means and standard deviations of all IPUI values combined for each condition (i.e. each replication within a condition was aggregated to get a single mean and standard deviation). Even though the difference between conditions after the item pool size 200 is not discernible in Figure 5.9, Table 5.2 shows that mean of IPUI increased steadiliy as the item pool size increased. Also the standard deviation of IPUI values also decreased (except for item pool size 40) as the item pool size increased. This shows that, item pools provided more appropriate items when item pool size increased.

Small standard deviations in Table 5.2 shows that smaller number of simulees suffer from the lack of appropriate item. This can be observed visually from Figure B.6 on page 179 which shows the distribution of IPUI for the 19th replication of each condition. This graph

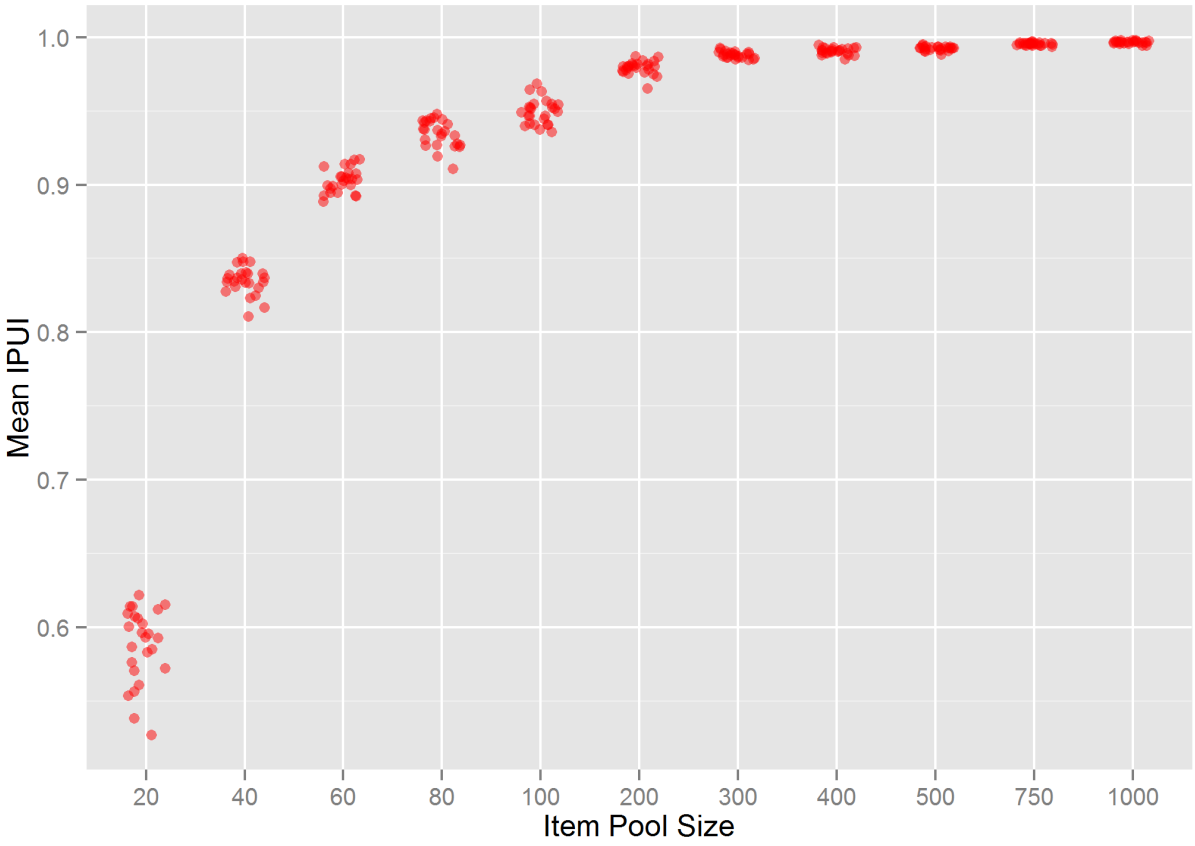

Figure 5.9: Mean IPUI Distribution by Item Pool Size Condition

Table 5.2: Means and Standard Deviations of IPUI Values by Item Pool Size

Item Pool Size	Mean IPUI	Standard Deviation of IPUI
20	0.5875	0.1798
40	0.8348	0.1817
60	0.9028	0.1593
80	0.9345	0.1325
100	0.9494	0.1201
200	0.9793	0.0768
300	0.9880	0.0574
400	0.9906	0.0507
500	0.9924	0.0450
750	0.9953	0.0343
1000	0.9965	0.0294

shows similar relationship between item pool size and IPUI as mentioned above. In addition to that, the number of examinees that have lower IPUI values decreased as the item pool size increased. This means, even the simulees with extreme ability parameters saw appropriate items. Figure B.6 also shows that the distribution of IPUI was negatively skewed for each condition.

IPUI and Standard Error Relationship Since both IPUI and SE uses information function for their calculations, these two values are expected to be related. Figure 5.10 shows the relationship between the mean IPUI and mean SE for each replication within a condition. Each item pool size condition is represented by a different color in the graph. It can be easily observed that there is almost a perfect relationship between mean IPUI and mean SE except for very small item pool size conditions. The correlation between these two variables was -0.99.

Figure 5.10 shows the relationship between IPUI and SE at an aggregate level. In fact, the relationship of IPUI and SE for individual simulees was not perfect. Figure 5.11 shows the relationship between IPUI and SE for the 19th replication of each condition. The relationship was more like a curvilinear relationship for individual examinees. This is very similar to the relationship observed in the first part of the results for Research Question 1 (Figure 5.4).

One interesting observation in Figure 5.11 is the correlation between SE and IPUI. As the size of the item pool increased the absolute value of correlation between SE and IPUI decreased. So, for item pool size 1000, the relationship between SE and IPUI appears to be very weak, $\rho = -0.34$. But for smaller item pools this relationship was rather strong.

Figure 5.12 shows the correlation between SE and IPUI for each replication. As the item pool size increased the absolute value of correlation between SE and IPUI decreased for most of the item pool size conditions. The main reason behind this is the variation of both SE and IPUI values. For smaller item pool sizes there was more variation in the values of these variables. But for larger item pool sizes the values of IPUI were almost 1, and the variation

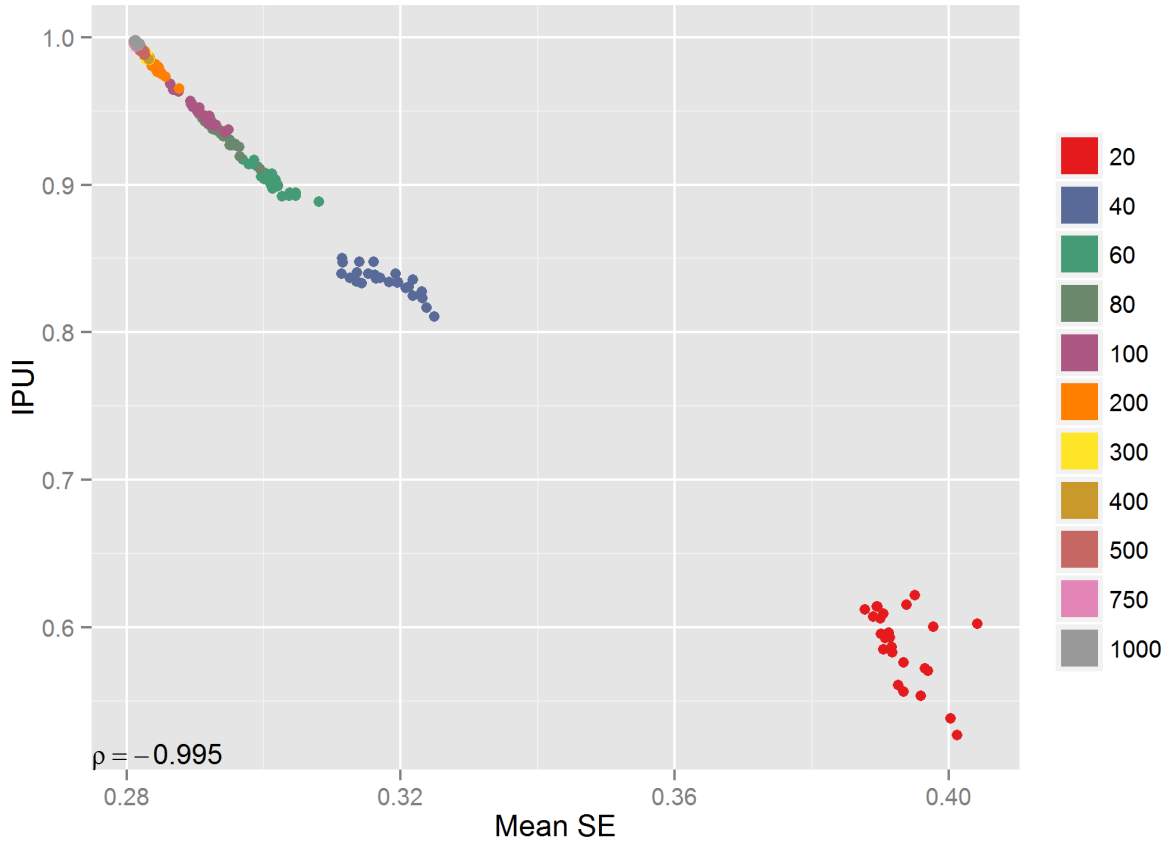

Figure 5.10: Relationship between Mean of IPUI and Mean of Standard Error

in SE was also very low and this was the main cause of weak relationship.

IPUI and Reliability Even though SE gives enough information about the error in ability estimates, reliability is a well known indicator of the quality of a test. It is valuable to investigate the relationship between reliability and IPUI.

The relationship between standard error (σ_e) and reliability ($\rho_{\theta,\hat{\theta}}$) is as follows:

$$\sigma_e = \sigma_{\hat{\theta}}\sqrt{1 - \rho_{\hat{\theta},\theta}}$$

So, reliability can be calculated as:

$$\rho_{\hat{\theta},\theta} = 1 - \frac{\sigma_e^2}{\sigma_{\hat{\theta}}^2}$$

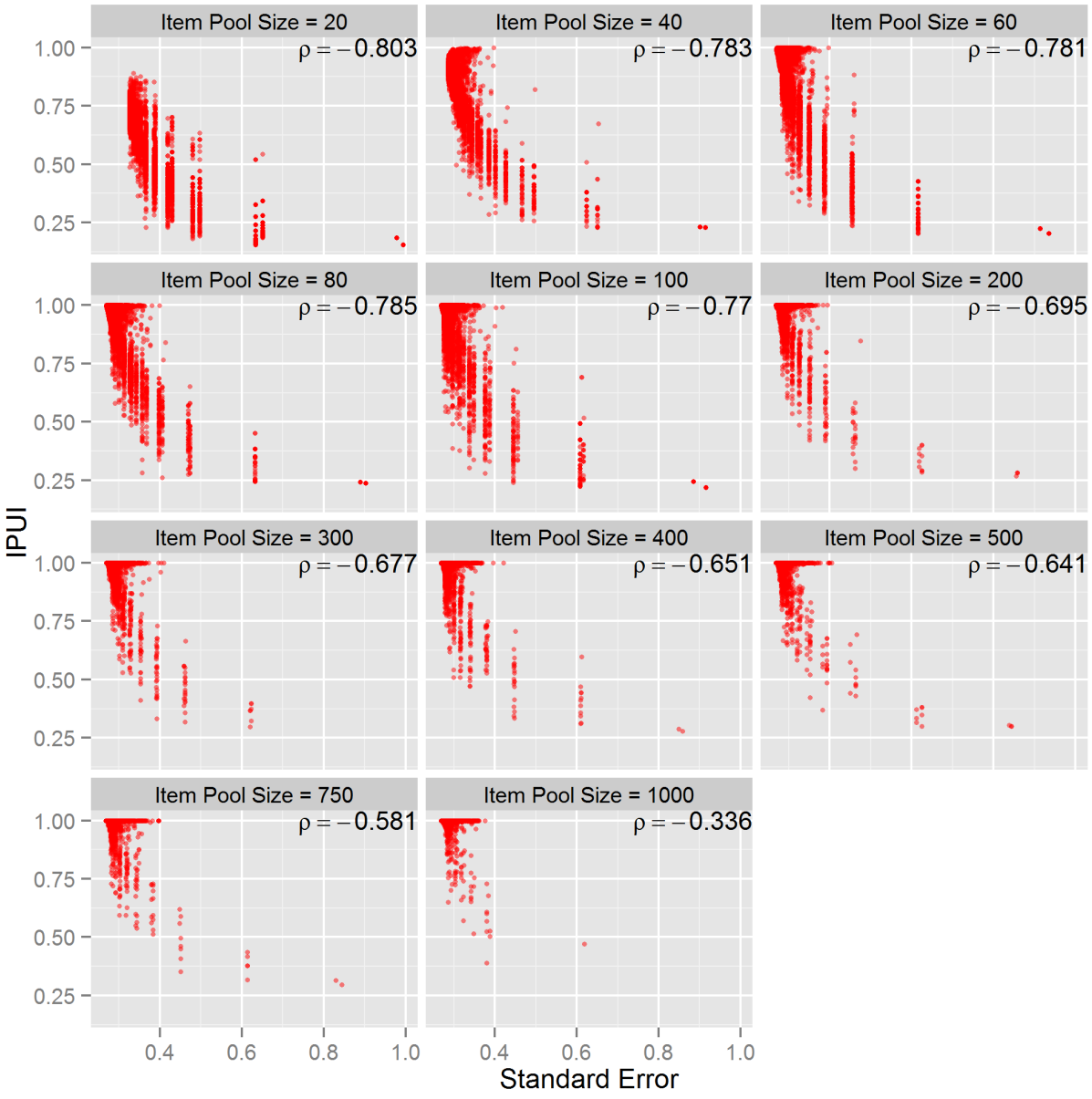

Figure 5.11: IPUI and Standard Error Relationship for Replication 19

In the simulation $\sigma^2_{\hat{\theta}} = 1$. Consequently, the reliability can be calculated as:

$$\rho_{\hat{\theta},\theta} = 1 - \sigma^2_e$$

Figure 5.13 shows the relationship between IPUI and reliability for the 19th replication of each condition. As the values of IPUI increased the reliability increased as well for each condition. The relationship was curvilinear instead of a linear one. For small item pools the

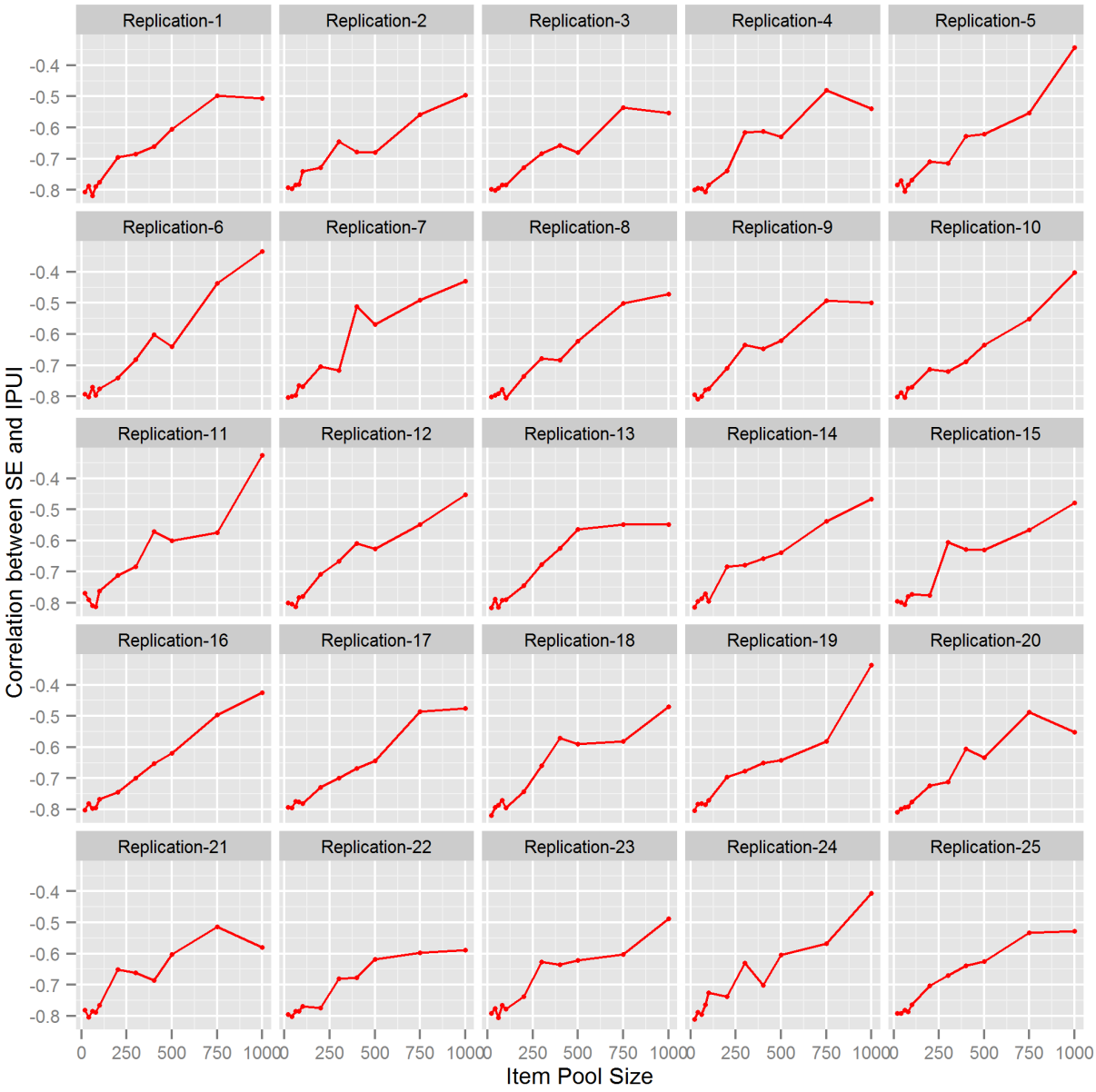

Figure 5.12: Correlation between Standard Error and IPUI for each Replication

relationship was rather strong. As item pool size increased the strength of the relationship decreased. Figure 5.13 also shows the mean reliability values for each condition (μ_x). The reliability was comparatively low for the item pool with 20 items. After the item pool size of 40, the increase was minimal.

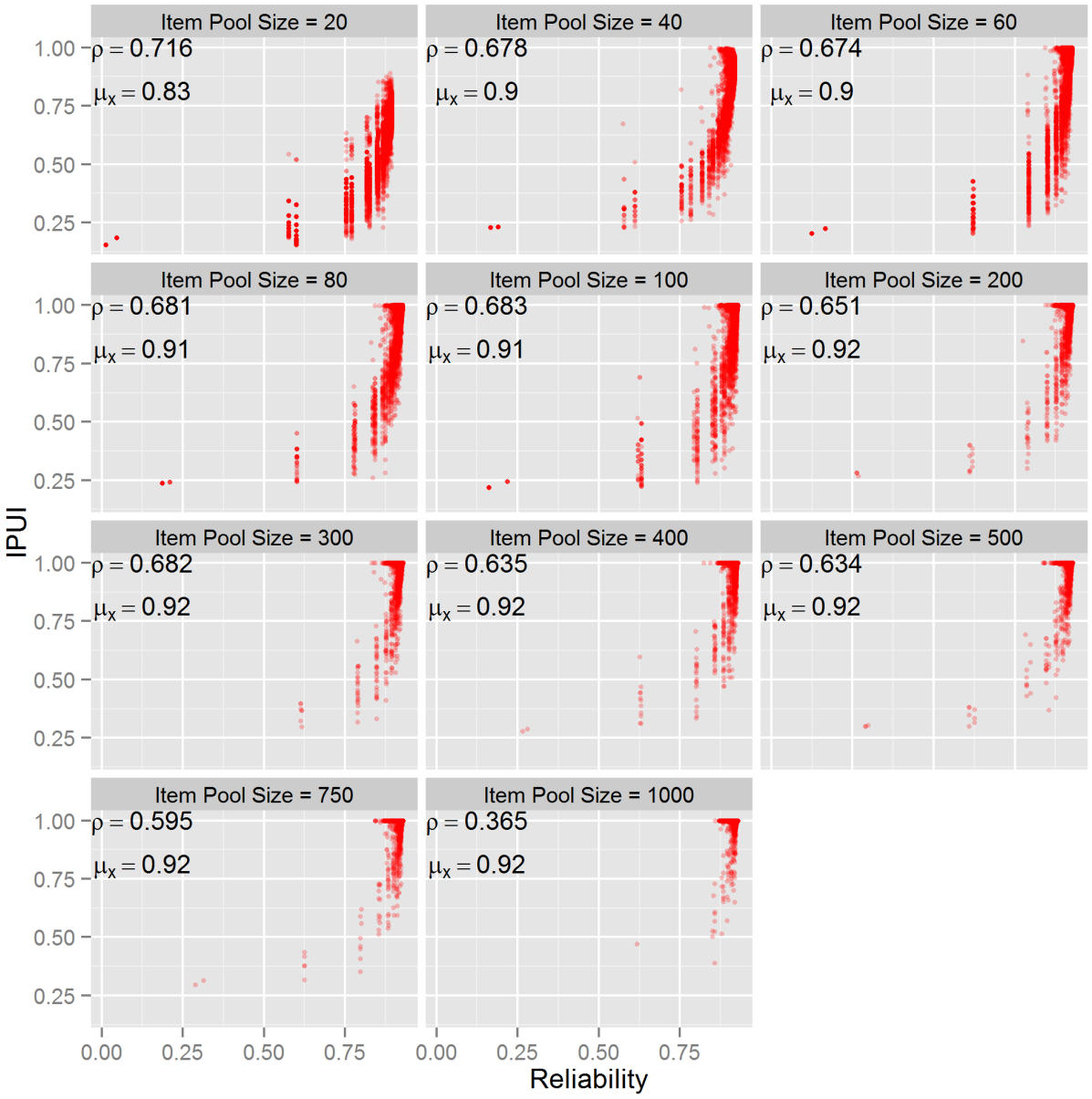

Figure 5.13: IPUI and Reliability Relationship for Replication 19

5.1.2 Research Question 2

In the Research Question 2, the effects of two test specifications, test length and exposure control, on the quality of item pool and CAT outcomes was investigated.

5.1.2.1 Test Length

In the first part of the Research Question 2, the effects of the test length on the performance of item pool and other CAT outcomes were investigated. It was hypothesized that increasing the test length will decrease the quality of the item pool as quantified by IPUI. Increasing the test length does not necessarily decrease the quality of a CAT, but the efficiency of a CAT will suffer if the item pool does not support a long test. For example, an increase in the test length is associated with an increase in the test's precision, i.e. decrease in the standard error of the ability estimates (Weiss, 1982). If the item pool is insufficient for a long test, increasing the test length will increase the precision of the test minimally. This hypothesis was tested using 18 test length conditions. The conditions were tests with lengths 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 100, 200, 300 and 400.

An item pool with 400 items was generated. Item difficulties (b-parameters) of this item pool were generated from a normal distribution with mean 0 and standard deviation 1. The item difficulty distribution of the item pool is shown in Figure C.1 on page 180. The same item pool was used for each test length condition. Since the size of the item pool was rather large, it was assumed that the effect of sampling error was low. So, different item pools did not replicated and only one replication was performed for each condition. 10,000 examinees were simulated from a normal distribution with mean 0 and standard deviation 1. The same set of examinees was used for each condition. The distribution of true θ is shown in Figure C.2 on page 181. There were no constraints on the item selection algorithm. The rest of the CAT specifications were the same as the common CAT specifications mentioned in Section 4.2.1.1 on page 37. For each condition, bias, SE, MSE, fidelity coefficient, exposure rates and IPUI values were calculated.

The results of the analyses are summarized in Figure 5.14. The fidelity coefficient and the mean values of bias, SE, MSE and IPUI for each test length condition are shown in the figure. Table 5.3 shows the fidelity coefficient in addition to the means and standard deviations of

bias, SE, MSE and IPUI for each test length condition. In the following pages the results of the investigation of each variable is presented separately.

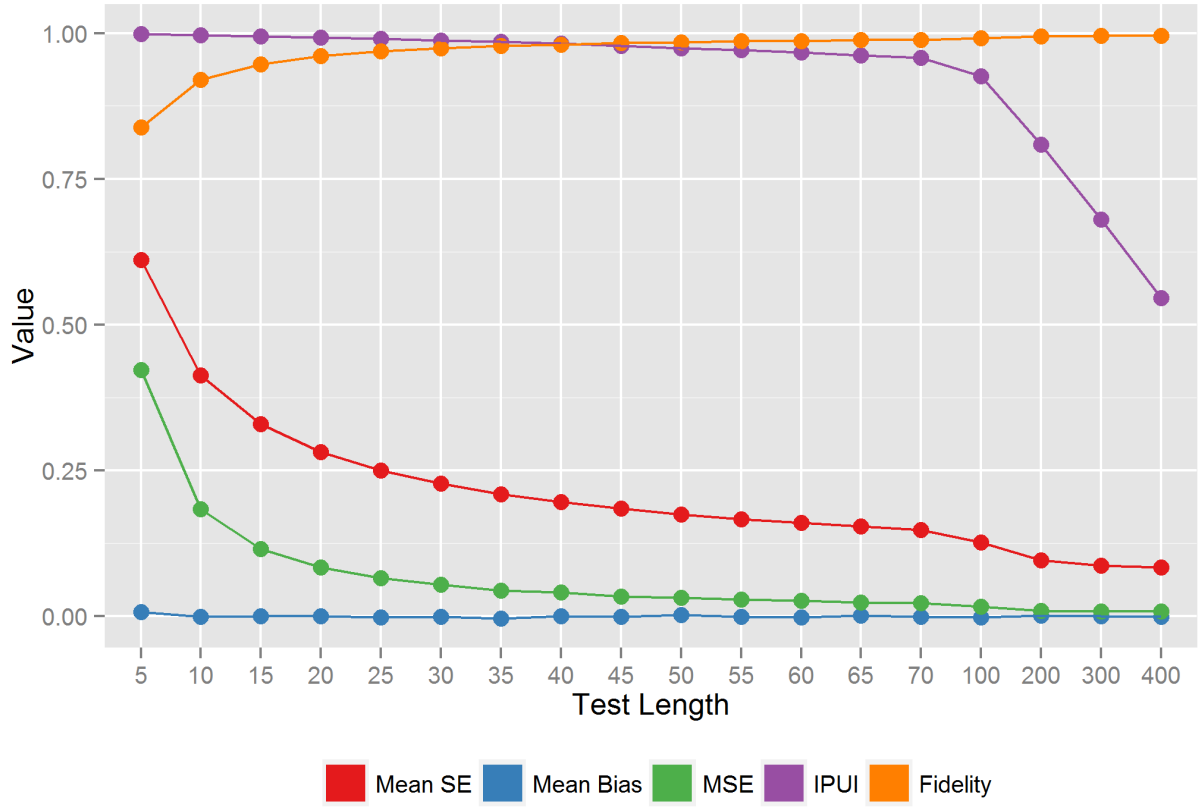

Figure 5.14: Summary Statistics for Research Question 2 - Test Length

Relationship between True and Estimated Ability Figure 5.15 shows the relationship between true ability (θ) and estimated ability ($\hat{\theta}$). The dashed lines in the figure are the identity lines (i.e. $y = x$ line). When the estimation is perfect, all of the points should fall on this line. If a point is above the identity line, this means it is overestimated, i.e. estimated ability is larger than the true ability (positive bias). If the point is below the identity line, this means it is underestimated, i.e. estimated ability is smaller than the true ability (negative bias).

At each condition, there was an error in the estimation and points deviated somewhat

Table 5.3: Summary Statistics for Research Question 2 - Test Length

Test Length	Bias	SE	MSE	IPUI	Fidelity
5	0.007 (0.650)	0.611 (0.046)	0.423 (0.653)	0.999 (0.004)	0.8390
10	-0.001 (0.429)	0.413 (0.030)	0.184 (0.289)	0.997 (0.021)	0.9198
15	0.000 (0.340)	0.330 (0.021)	0.116 (0.172)	0.995 (0.032)	0.9473
20	0.000 (0.289)	0.282 (0.017)	0.084 (0.122)	0.992 (0.042)	0.9609
25	-0.002 (0.256)	0.250 (0.015)	0.065 (0.099)	0.991 (0.049)	0.9690
30	-0.001 (0.234)	0.227 (0.016)	0.055 (0.092)	0.988 (0.055)	0.9740
35	-0.004 (0.211)	0.210 (0.013)	0.044 (0.065)	0.985 (0.059)	0.9787
40	-0.000 (0.202)	0.196 (0.016)	0.041 (0.063)	0.982 (0.067)	0.9806
45	-0.000 (0.185)	0.185 (0.015)	0.034 (0.052)	0.979 (0.073)	0.9834
50	0.002 (0.179)	0.175 (0.014)	0.032 (0.048)	0.975 (0.081)	0.9845
55	-0.001 (0.169)	0.167 (0.015)	0.028 (0.042)	0.971 (0.084)	0.9862
60	-0.002 (0.164)	0.160 (0.017)	0.027 (0.047)	0.967 (0.089)	0.9869
65	0.002 (0.155)	0.154 (0.015)	0.024 (0.039)	0.962 (0.094)	0.9883
70	-0.001 (0.150)	0.148 (0.014)	0.022 (0.034)	0.958 (0.097)	0.9891
100	-0.002 (0.128)	0.126 (0.017)	0.016 (0.031)	0.926 (0.123)	0.9920
200	0.002 (0.099)	0.096 (0.020)	0.010 (0.023)	0.809 (0.164)	0.9952
300	0.001 (0.090)	0.087 (0.020)	0.008 (0.015)	0.681 (0.167)	0.9960
400	-0.001 (0.089)	0.084 (0.023)	0.008 (0.026)	0.546 (0.143)	0.9961

Note. Numbers within the parentheses are standard deviations of each outcome.
SE: Standard Error; MSE: Mean Squared Error; Fidelity: Correlation between true ability and estimated ability

from the identity line. For short tests the deviation from the identity line is larger. As the length of the test increased, the points get closer to the identity line. This means the estimation started to converge to the true values. Both Figure 5.15 and Table 5.3 shows the fidelity coefficients (i.e. the correlation between true and estimated ability) for each test length condition. For shorter tests the correlation between true and estimated abilities were lower. As test length increased, the correlations steadily increased and approached to 1.

Bias Figure 5.14 shows that the mean bias did not change across different conditions. On the other hand, it can be observed from Table 5.3 that, the standard deviation of bias decreases as the test length increases. Figure 5.16 shows the decrease in the standard deviation of bias values visually. When the test length was short, the error in the estimates were higher

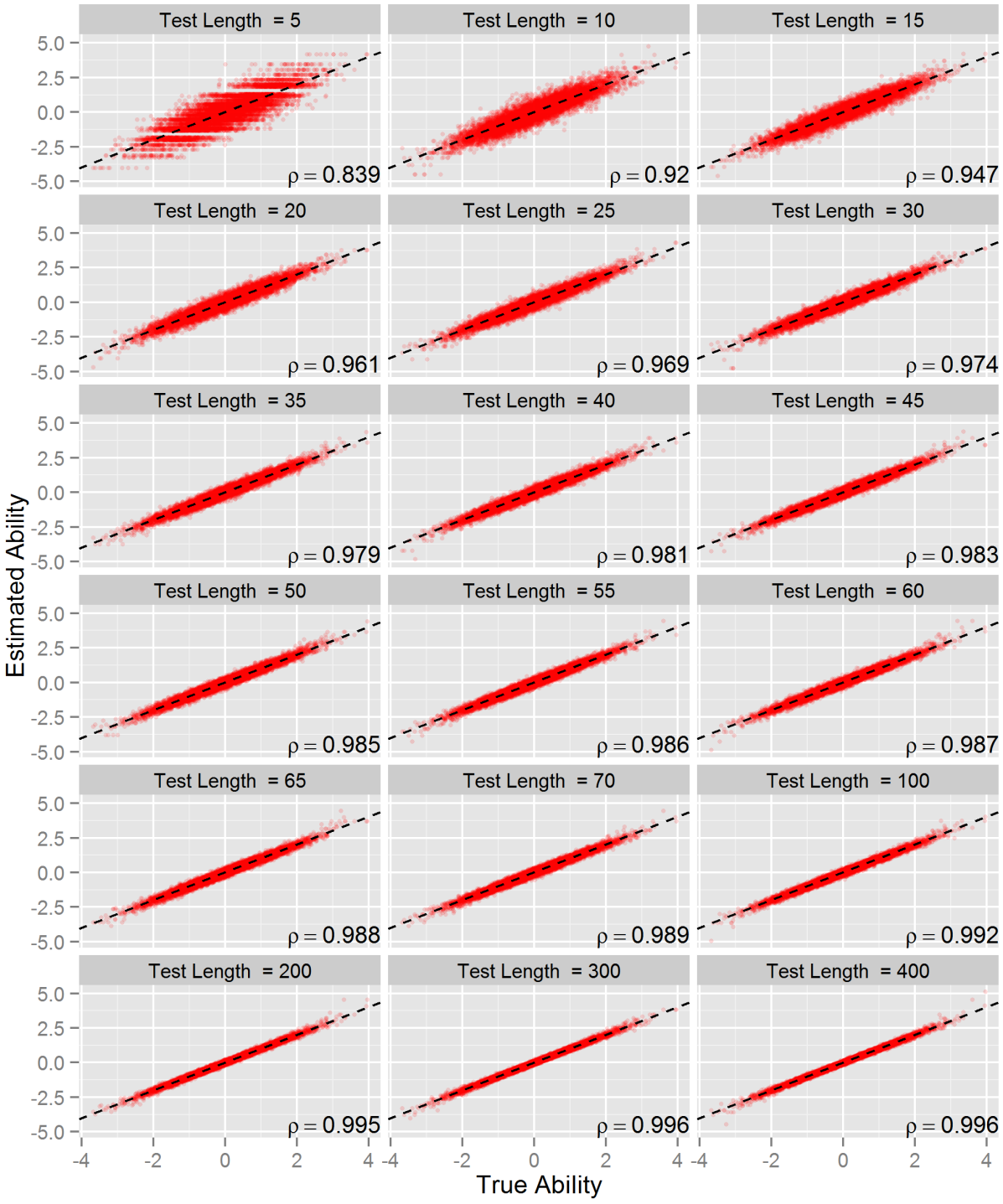

Figure 5.15: Relationship between True and Estimated Ability by Test Length Condition

as also shown in Figure 5.15. This is not surprising because longer tests are expected to increase the quality of the measurement.

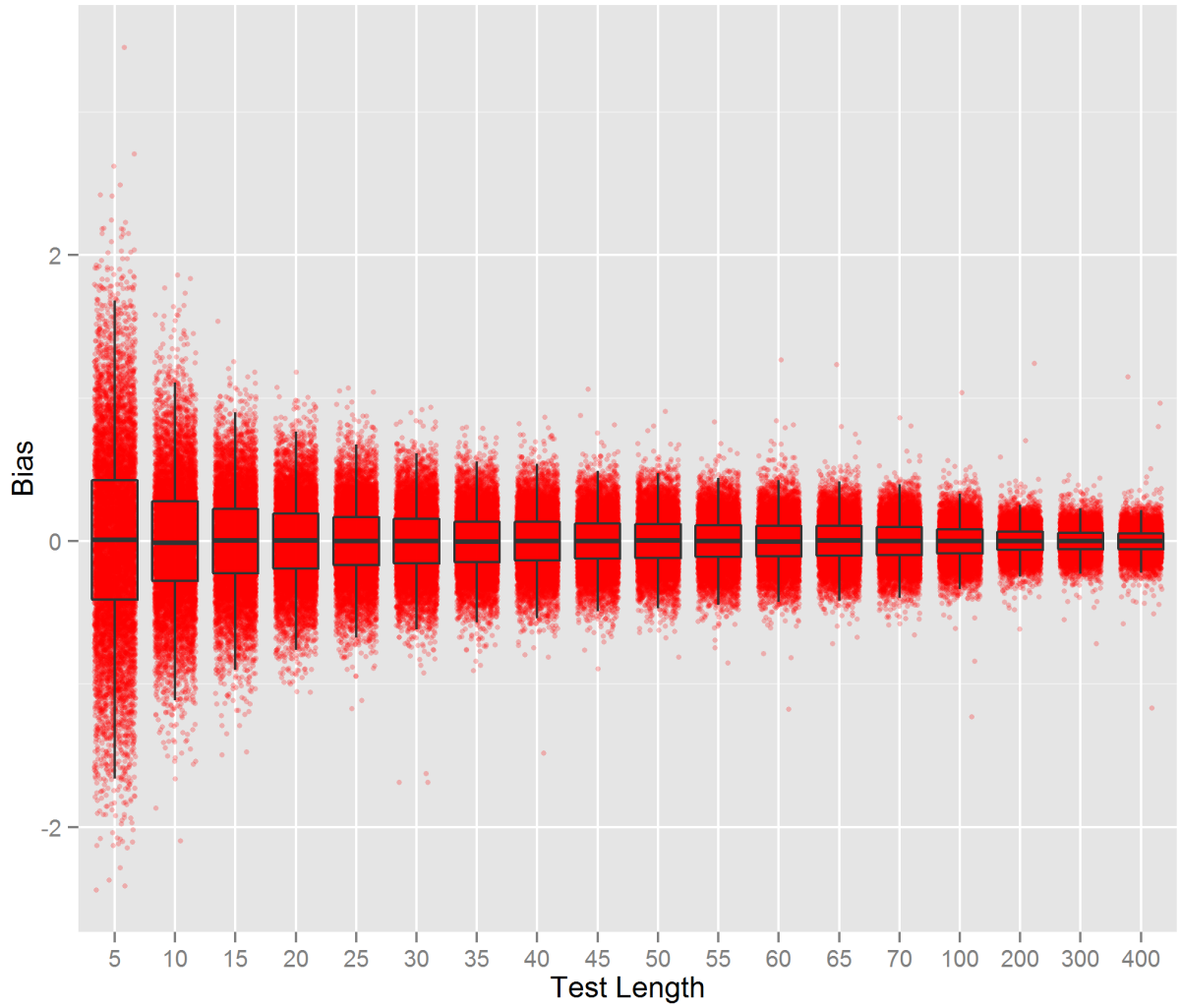

Figure 5.16: Bias Distribution by Test Length Condition

Standard Error It can be observed from Figure 5.14 that as the length of test increased the mean value of SE decreased. Table 5.3 shows the standard deviations of SEs for each test length condition. These values does not show a uniform pattern.

The distribution of SEs can be observed visually at Figure 5.17. In this figure, the dashed line corresponds to the standard error that is expected for a test with reliability value 0.9. For test length conditions longer than 20, majority of the simulees had SE values lower than this threshold. The median of SEs decreased quickly between test length 5 to 30. After this there was a decrese but this decrease was not large. Especially, there was a minimal decrease

from test length 300 to 400, even though the test length increased by 100 items. This is an evidence for the efficiency loss for long tests in glscat.

At each test length condition, the distributions of SEs had a strong positive skew. The reason behind this was the overlap between the ability distribution of examinees and item difficulty distribution of item pool. For most of the examinees, the item pool had appropriate items. For a small portion of examinees who were at the extremes of the ability scale, the item pool did not have appropriate items. The SEs of these examinees were larger compared to the examinees that were closer to the center of the ability distribution. Since the number of examinees at the extremes were small compared to the ones at the center, a positive skew occurred for SE distribution.

Mean Squared Error The average of MSE values decreased as the test length increased (Figure 5.14). This decrease was rapid for shorter test lengths but the trend was flattened after the test length 100. Table 5.3 shows that the standard deviation of MSE values also decreased as the test length increased. Figure 5.18 shows the distribution of MSE for each test length condition. The spread of MSE values decreased as the test length increased. For test lengths longer than 100, there was almost no difference between the MSE distributions.

Exposure Rates Exposure rate distribution of each test length condition is shown in Figure 5.19. In this figure, median exposure rates are shown with bold lines in the middle of box-plots fitted to the exposure rates of each condition. The dashed lines are showing the 0.20 and 0.05 levels for exposure rates. It is recommended that exposure rates of the items fall between these two dashed lines. For very small test length conditions the exposure rates were very low. Exposure rates of the items increased as test length increased, because the item pool size was the same for each test length condition.

Table 5.4 shows the means and standard deviations of the exposure rates in addition to the rates of exposed items that were larger than 0.20 and lower than 0.05. An exposure rate

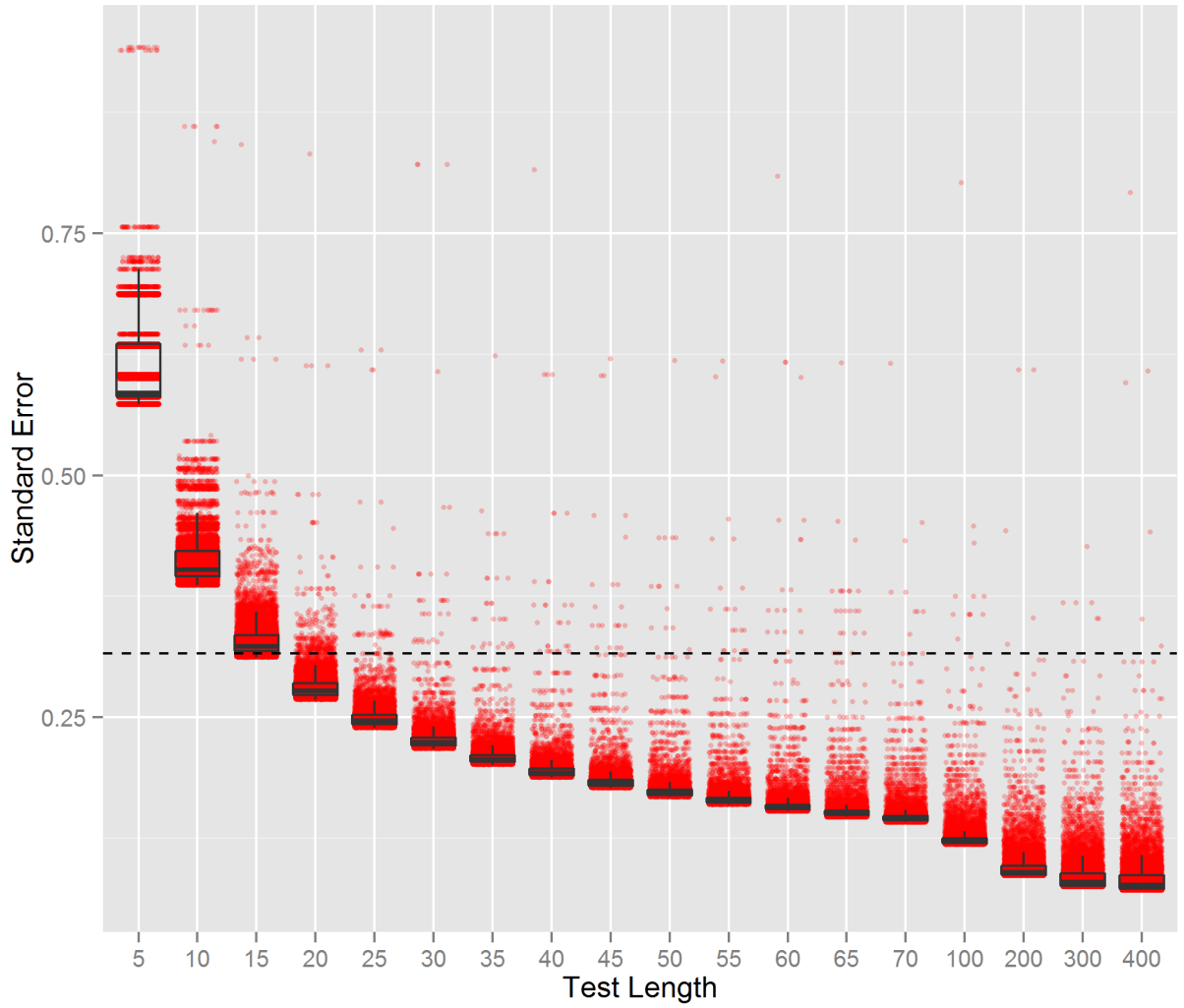

Figure 5.17: Standard Error Distribution by Test Length Condition

larger than 0.20 can be seen as a sign of a highly exposed item. It is desirable for a CAT test to have item exposures lower than 0.2. On the other hand, very low exposure rates signify items that are not shown to the majority of the examinees. This is not desirable because underutilization of items decreases the efficiency of the item pools. In this sense 0.05 value can be seen as a cutoff value for items with low exposure rates. These cutoffs can change depending on the purpose of the test.

For test length conditions lower than 35, almost none of the items were overexposed (i.e. exposure rates > 0.2). The small percentage of items ($400 \times 0.0175 = 7$ items) that had larger

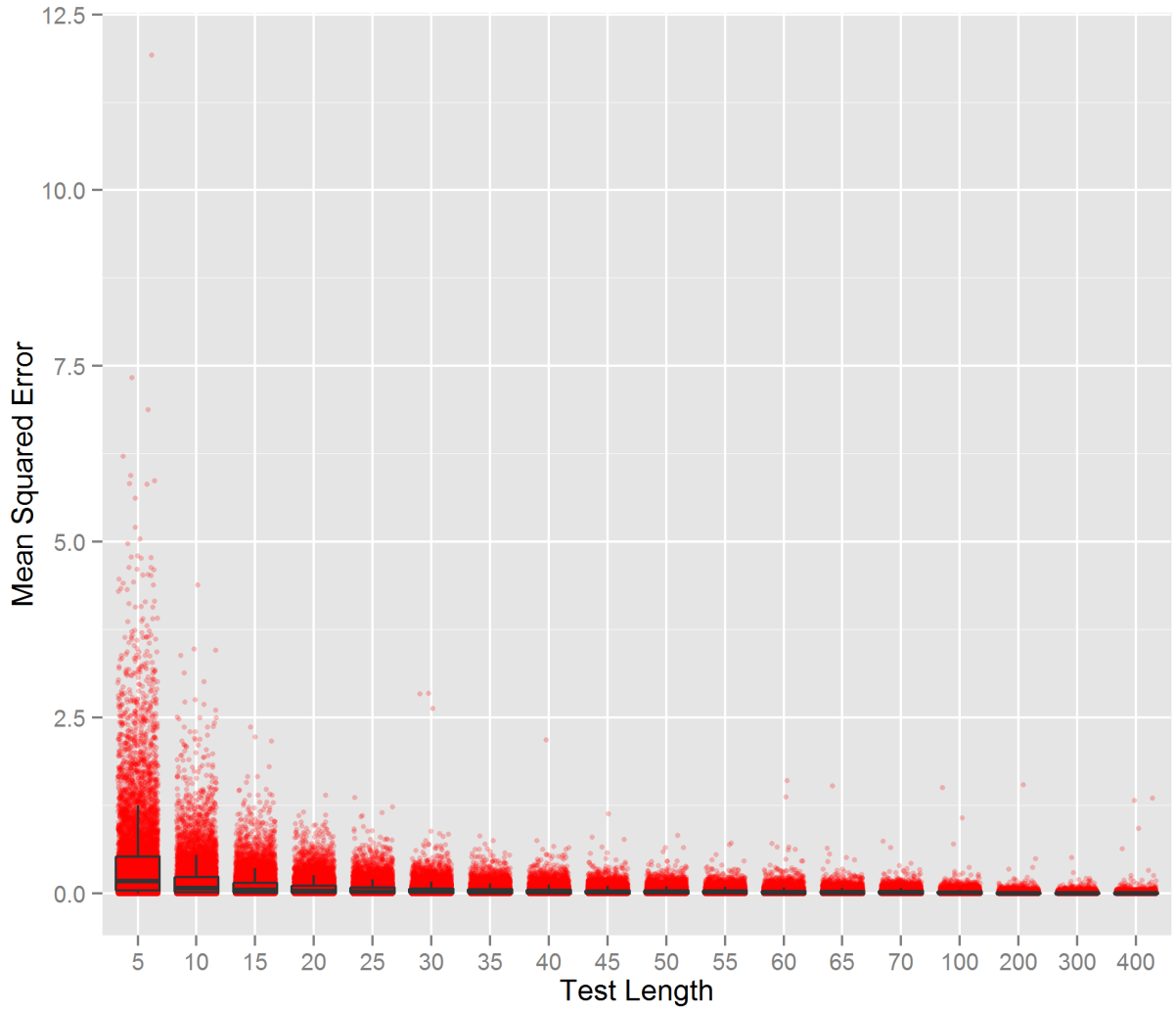

Figure 5.18: Mean Squared Error Distribution by Test Length Condition

exposure rates was due to the lack of exposure control in the item selection algorithm, as explained previously in Section 5.1.1.2. For test length conditions longer than 55 items more than 10% of the items were overexposed. The proportion of underexposed items decreased as test length increased. For test length conditions shorter than 45 items, more than 10% of the items were underexposed.

IPUI Figure 5.14 shows that the mean value of IPUI decreased as the test length increased. Standard deviations of IPUI values increased as test lengths increased as shown in Table 5.3.

Figure 5.19: Item Exposure Distribution by Test Length Condition

Clearly, the item pool was able to support shorter tests. Mean IPUI value for tests with 5 items were almost perfect. Also low standard deviations for shorter tests point out that most of the simulees saw appropriate items. Figure 5.20 shows the IPUI distribution of simulees visually. The number of examinees with low IPUI values increased as the test length increased. For test lengths larger than 100, none of the examinees had IPUI values larger than .98, and this maximum value decreased as test length increased.

In Figure 5.20, each test length condition shows a negative skew. The reason behind this skewness is the overlap between the item difficulty distribution of the item pool and the ability distribution of the simulees. For the majority of the simulees, item pool had items that were close to their true θ values. Consequently, the IPUI values were high for most of the simulees. The IPUI values of the simulees with true θ values at the extremes were lower.

Table 5.4: Item Exposure Analysis by Test Length Condition

Test Length	Mean Exposure	SD	Exposure > .20	Exposure < .05
5	0.0125	0.0709	0.0175	0.9450
10	0.0250	0.0712	0.0175	0.9050
15	0.0375	0.0707	0.0175	0.8200
20	0.0500	0.0704	0.0175	0.6500
25	0.0625	0.0704	0.0175	0.4750
30	0.0750	0.0704	0.0175	0.3275
35	0.0875	0.0702	0.0200	0.2150
40	0.1000	0.0695	0.0275	0.1050
45	0.1125	0.0692	0.0325	0.0500
50	0.1250	0.0690	0.0425	0.0250
55	0.1375	0.0692	0.0675	0.0250
60	0.1500	0.0694	0.1125	0.0200
65	0.1625	0.0702	0.1875	0.0175
70	0.1750	0.0704	0.2550	0.0175
100	0.2500	0.0822	0.8425	0.0125
200	0.5000	0.2041	0.9350	0.0050
300	0.7500	0.2836	0.9725	0.0000
400	1.0000	0.0000	1.0000	0.0000

But since the number of the simulees with extreme true θ values were low, the distribution of the IPUI became skewed. The amount of skewness decreased as the test length increased. For longer tests, item pool failed to provide appropriate items even for the examinees that had true θ values close to 0. The IPUI values of all examinees started to decrease. Since IPUI has a lower bound of 0, and unlike the upper bound of 1, this lower bound cannot be attainable practically, as IPUI decreased for all simulees, the IPUI values of the simulees with true θs at the middle of the ability distribution decreased more compared to the simulees at the extremes. This is the reason why the skewness of IPUI distribution decreased for longer test length conditions.

One important difference between Figure 5.14 and Figure 5.20 is the use of summary statistics to show the difference between IPUI values. In Figure 5.14 the mean values of IPUI are shown. On the other hand, on Figure 5.20 the median values of IPUI are shown in addition to the quartile information via box plots. Since IPUI distributions were negatively

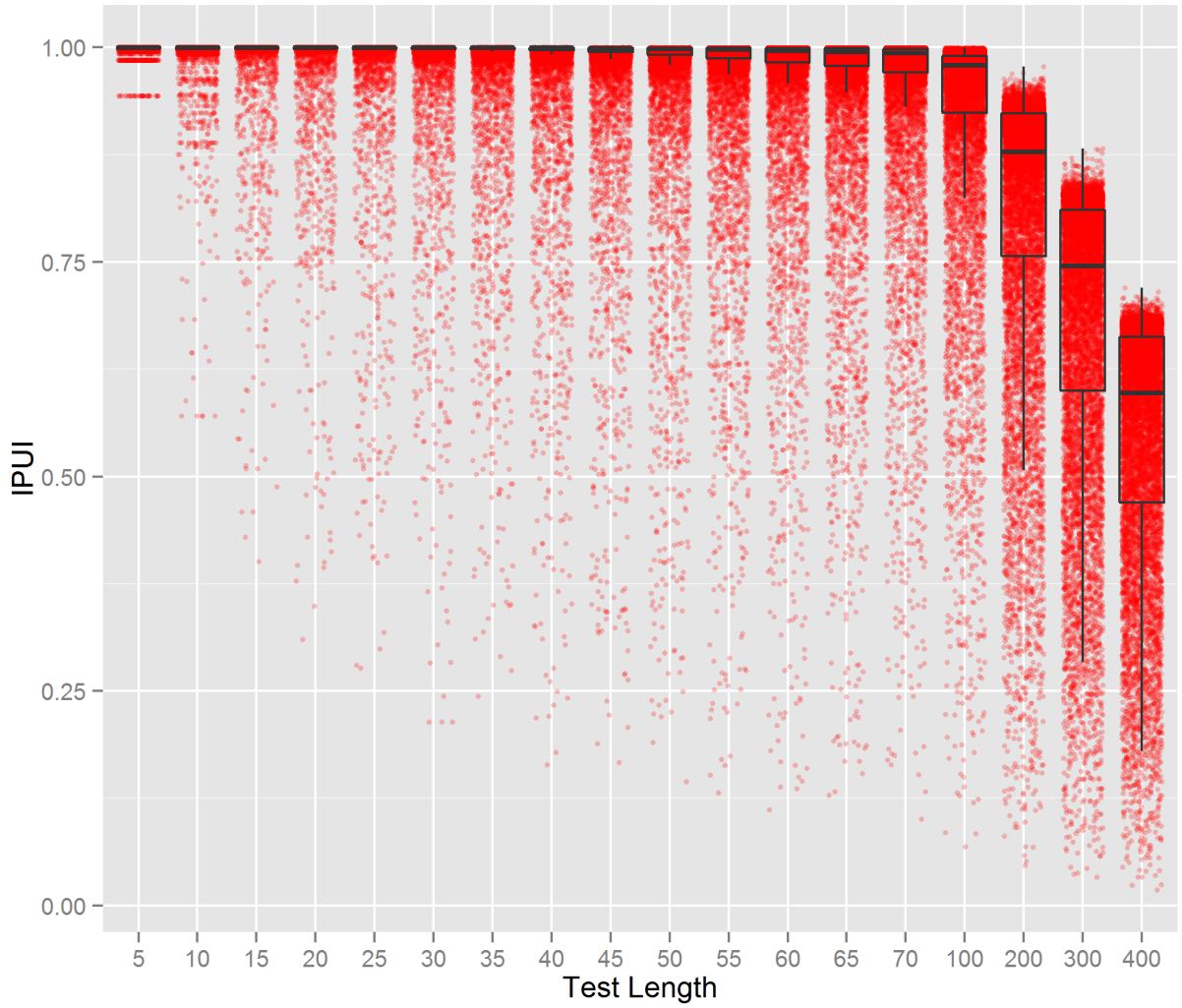

Figure 5.20: IPUI Distribution by Test Length Condition

skewed for all conditions, the median of IPUI values were always lower than the mean values. This difference might have important consequences on how to evaluate the overall IPUI values. For example, for the simulation condition where test length was 70, the mean and median of IPUI values were 0.958 and 0.994, respectively. According to median value the item pool function almost perfectly, on the other hand mean of IPUI values indicates that item pool performance was not perfect.

IPUI and Standard Error In the previous sections the relationship between mean values of IPUI and SE were negative. As the discrepancy between item pool and ability distribution increased in the first part of the Research Question 1, the mean of IPUI values decreased and the mean of SE valued increased. The same type of relationship was observed for the second part of the Research Question 1 where the item pool size investigated. As the size of the item pool increased, the mean of IPUI values increased and the mean of SE valued decreased. For test length conditions, this trend changed. Figure 5.14 on page 74 shows that as test length increased, the mean values of both IPUI and SE decreased. In this case, even though increasing the test length decreased the quality of item pool for that particular CAT design, the precision of the ability estimates increased. This is an important observation that shows how different CAT specifications interact with the outcomes of CAT differently.

The relationship between IPUI and SE for each test length condition is shown in Figure 5.21. For individual simulees, the relationship between SE and IPUI is negative, similar to what was observed in previous sections (Figures 5.4 and 5.11 on page 59 and on page 70). Different than the previous sections, the curvilinear relationship between these two variables became more evident as the test length increased. Figure 5.21 also shows the correlations between SE and IPUI for each test length condition. The correlation between SE and IPUI was larger in absolute sense for longer tests.

5.1.2.2 Exposure Control

For the second part of the Research Question 2, the effects of the exposure control on the performance of the item pool and other CAT outcomes was investigated. It was hypothesized that adding more randomization to the item selection process to reduce item exposure will decrease the quality of the item pool as quantified by IPUI. This hypothesis was tested using 12 exposure control conditions. The conditions were (1) no exposure control, (2-11) randomesque with 3, 5, 7, 10, 13, 15, 20, 25, 50, 100 items and (12) total-random-selection of items. These exposure control conditions started from the condition where there was no

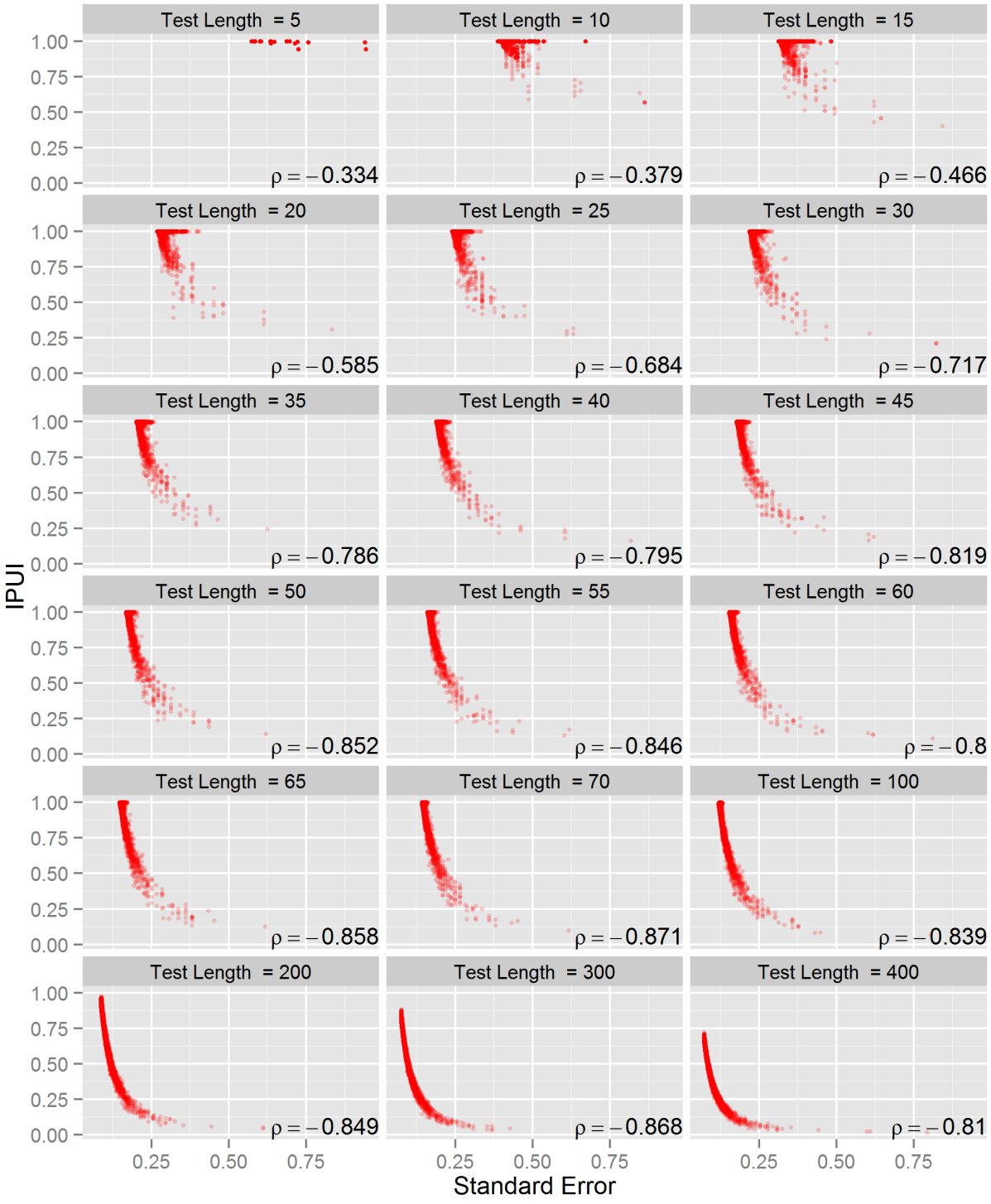

Figure 5.21: IPUI and Standard Error Relationship by Test Length Condition

randomization in the item selection and gradually increased the amount of randomization imposed on the item selection algorithm. The last condition basically randomly administered items to examinees without taking into account their previous answers. In the following pages these conditions will be referred as: Rand-1, Rand-3, Rand-5, Rand-7, Rand-10, Rand-13, Rand-15, Rand-20, Rand-25, Rand-50, Rand-100 and Rand-Total.

An item pool with 250 items was generated. Item difficulties (b-parameters) of this item pool were generated from a normal distribution with mean 0 and standard deviation 0.7. The item difficulty distribution of the item pool is shown in Figure D.1 on page 183. The same item pool was used for each exposure control condition. 10,000 examinees were simulated from a normal distribution with mean 0 and standard deviation 1. The same set of examinees was used for each condition. The distribution of true θ is shown in Figure D.2 on page 184. Test length for all of the tests was 20. There were no constraints on the item selection algorithm other than exposure control. The rest of the CAT specifications were the same as the common CAT specifications mentioned in Section 4.2.1.1 on page 37. For each condition, bias, SE, MSE, fidelity coefficient, exposure rates and IPUI values calculated.

The results of the analyses are summarized in Figure 5.22. The mean values of bias, SE, MSE, IPUI and fidelity coefficient for each test length condition is shown in the figure. Table 5.5 shows the means and standard deviations of biases, SEs, MSEs and IPUIs for each test length condition in addition to the fidelity coefficients. In following pages each of these output variables are discussed separately.

Relationship between True and Estimated Ability Figure 5.23 shows the relationship between true ability (θ) and estimated ability ($\hat{\theta}$). The dashed lines in the figure are the identity lines (i.e. $y = x$ line). At each condition, there was an error in the estimation and points deviated somewhat from the identity line. Towards the middle of the ability scale, there is a balance between overestimated and underestimated estimates. But towards the extremes of the ability scale, balance starts to shift. Ability was overestimated for high ability

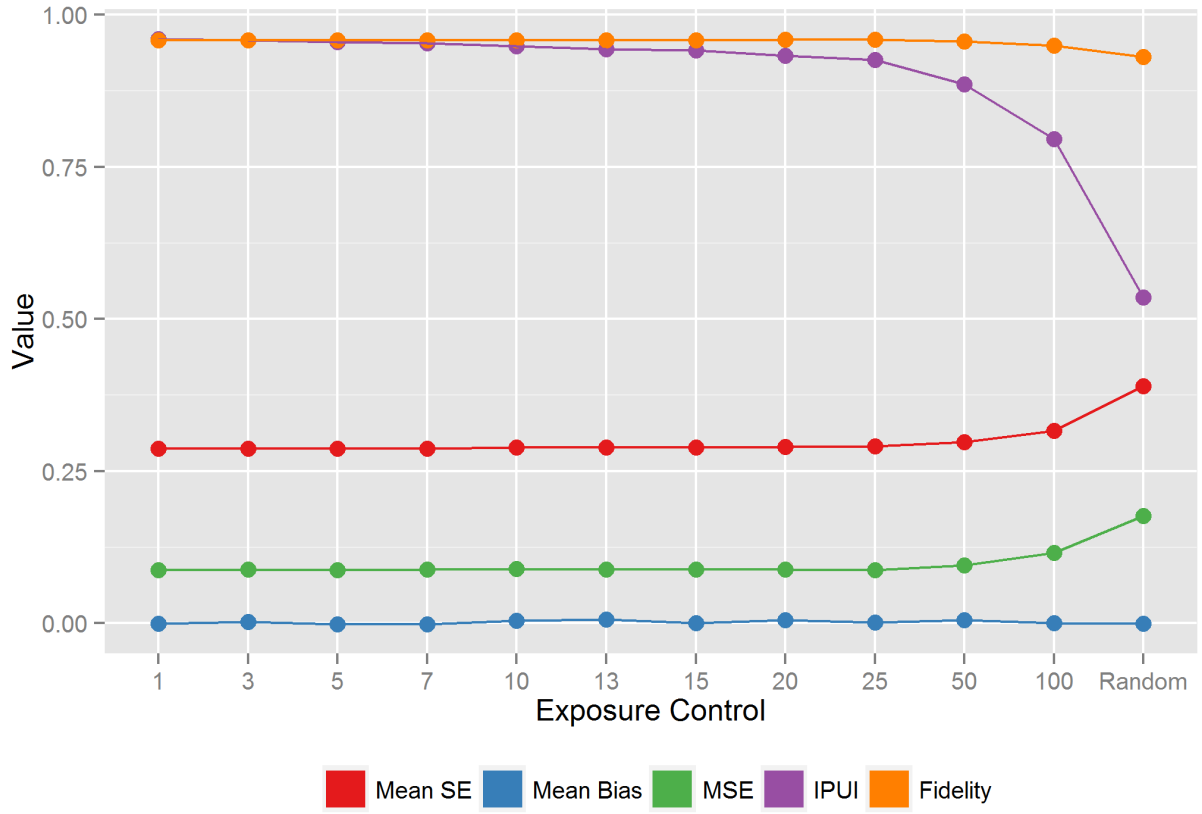

Figure 5.22: Summary Statistics for Research Question 2 - Exposure Control

examinees and underestimated for low ability examinees.

Figure 5.23 also shows the fidelity coefficients (i.e. correlation between true and estimated ability) for each exposure control condition. For conditions Rand-1 to Rand-25, the fidelity coefficient was not changed between conditions. But after Rand-25, fidelity coefficient started to decrease.

Bias Figure 5.22 shows that mean bias changes minimally across different conditions. This minimal change was not in one direction, so it can be treated as a random error. Table 5.5 also shows that the magnitude of these mean biases were almost 0. The standard deviation of biases were almost the same for conditions in which randomization was less than 25 items. After Rand-25 condition the standard deviations of the biases started to increase

Table 5.5: Summary Statistics for Research Question 2 - Exposure Control

Condition	Bias	SE	MSE	IPUI	Fidelity
1	-0.000 (0.295)	0.287 (0.044)	0.087 (0.149)	0.961 (0.113)	0.9580
3	0.002 (0.297)	0.287 (0.046)	0.088 (0.150)	0.958 (0.117)	0.9582
5	-0.001 (0.296)	0.287 (0.046)	0.088 (0.151)	0.956 (0.118)	0.9586
7	-0.002 (0.297)	0.288 (0.047)	0.088 (0.151)	0.953 (0.120)	0.9581
10	0.004 (0.299)	0.289 (0.053)	0.089 (0.166)	0.949 (0.127)	0.9581
13	0.006 (0.297)	0.289 (0.053)	0.089 (0.156)	0.944 (0.130)	0.9587
15	0.000 (0.297)	0.289 (0.051)	0.088 (0.163)	0.941 (0.130)	0.9586
20	0.005 (0.297)	0.290 (0.052)	0.088 (0.148)	0.933 (0.135)	0.9588
25	0.001 (0.296)	0.291 (0.057)	0.088 (0.156)	0.926 (0.140)	0.9589
50	0.005 (0.309)	0.298 (0.072)	0.095 (0.187)	0.886 (0.163)	0.9563
100	0.001 (0.341)	0.317 (0.101)	0.116 (0.249)	0.796 (0.188)	0.9494
Random	-0.000 (0.420)	0.389 (0.150)	0.176 (0.375)	0.535 (0.180)	0.9307

Note. Numbers within the parentheses are standard deviations of each outcome.
Condition: Exposure Control; SE: Standard Error; MSE: Mean Squared Error; Fidelity: Correlation between true ability and estimated ability

systematically. This can also be observed visually from Figure 5.24. The fact that mean biases did not change makes sense because (1) item pool and ability distribution is overlapping and (2) both high and low ability examinees were affected by randomization. From the standard deviations of biases it can be said that the item pool can support randomization till Rand-25. But after that, the amount of randomization started to affect the ability estimates of examinees.

Standard Error As Figure 5.22 shows, the increase in the mean SE was almost nonexistent from the condition Rand-1 to Rand-25. After Rand-25, there was an increase in the mean SE values. Table 5.5 also shows an increase but very minimal for low randomization conditions. In addition, Table 5.5 shows that standard deviation of SEs increased steadily as the amount of randomization in item selection increased. Visual inspection of the spread at Figure 5.25 indicates this as well. In Figure 5.25 the dashed line points the SE value for a test with 0.9 reliability as explained at Section 5.1.1.2 on page 69. Majority of the simulees had SEs smaller than 0.316 for conditions Rand-1 to Rand-50. The upper whisker of the box-plots

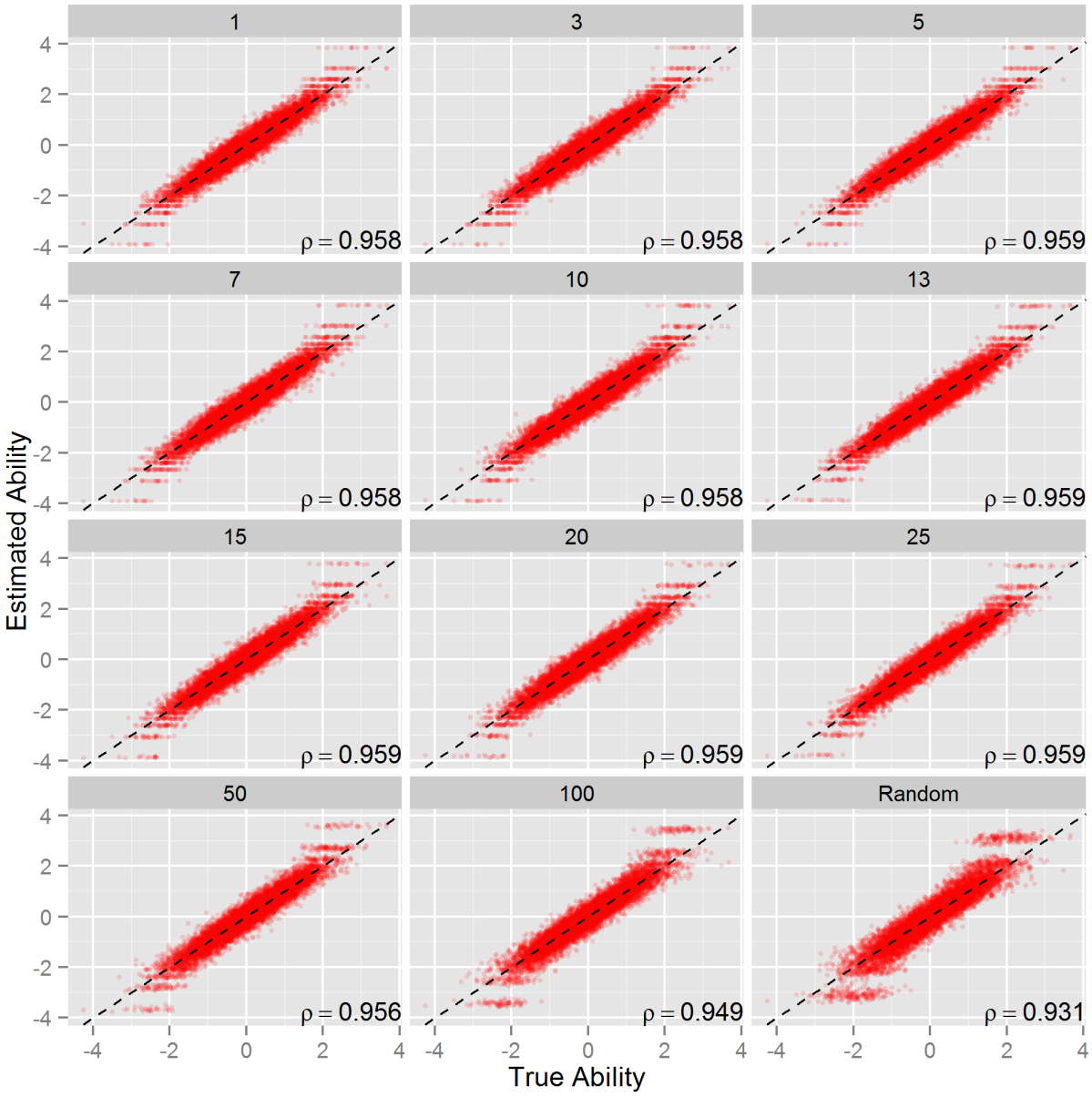

Figure 5.23: Relationship between True and Estimated Ability by Exposure Control Condition

are lower than 0.316. For total random selection condition, most of the simulees had SEs larger than this threshold. For Rand-100 and Rand-Total conditions, the negative effects of randomization on ability estimates are more visible.

Mean Squared Error Figure 5.22 shows that the pattern in average MSE values were almost identical to the pattern of SEs. The numbers in Table 5.5 indicates that for conditions

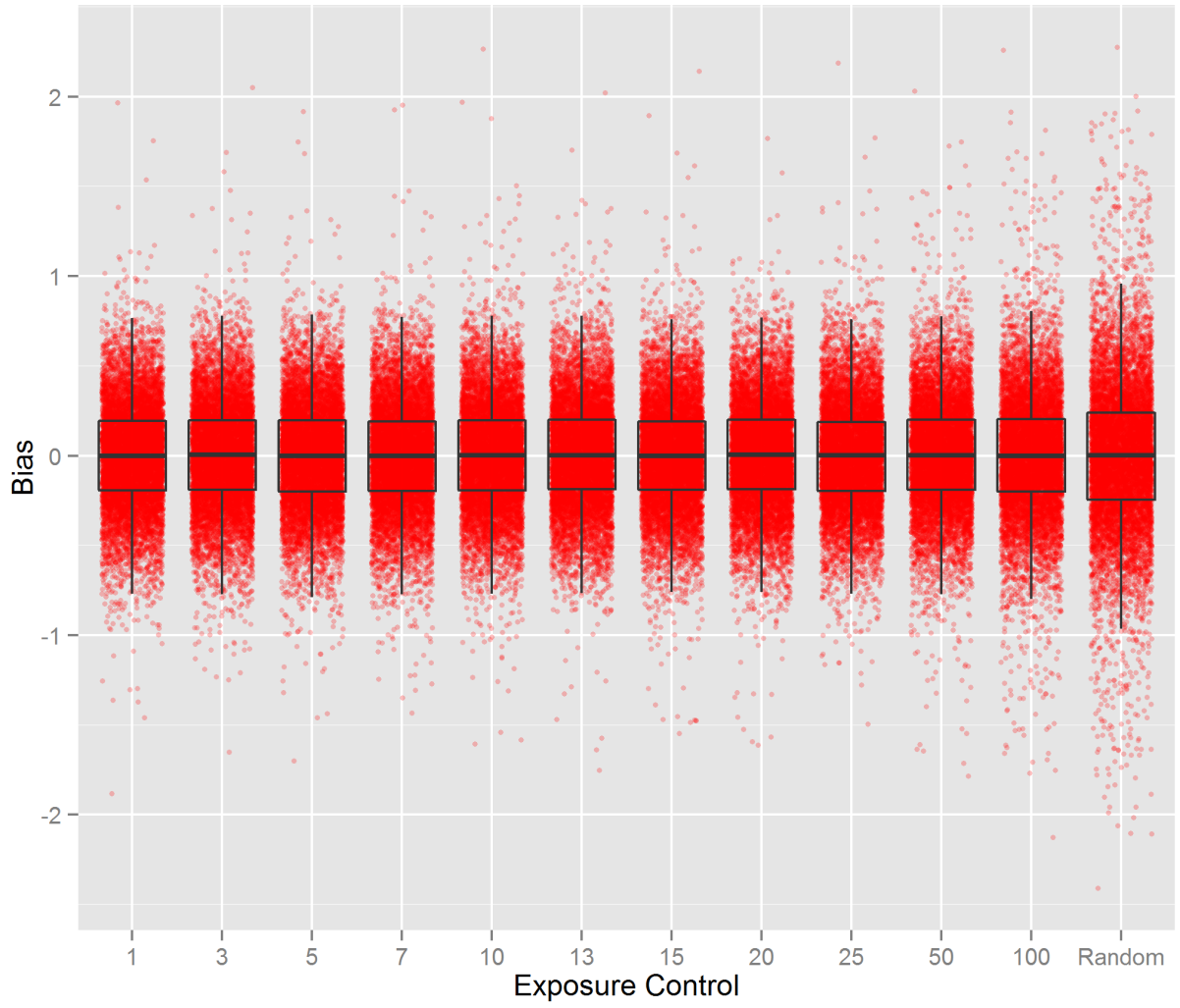

Figure 5.24: Bias Distribution by Exposure Control Condition

between Rand-1 and Rand-25, the change in the average MSE values was minimal and not steady. But there was an increasing pattern for conditions from Rand-50 to Rand-Total. The change in the standard deviations of MSE values was not steady between Rand-1 and Rand-25 conditions. For example, standard deviation of MSE for Rand-20 condition was less than the standard deviation for the Rand-10 condition. On the other hand, the increase in the standard deviation of MSE values is evident for conditions Rand-50 to Rand-Total. Figure 5.26 shows this spread visually. Especially for Rand-Total condition, the increase in the spread is evident.

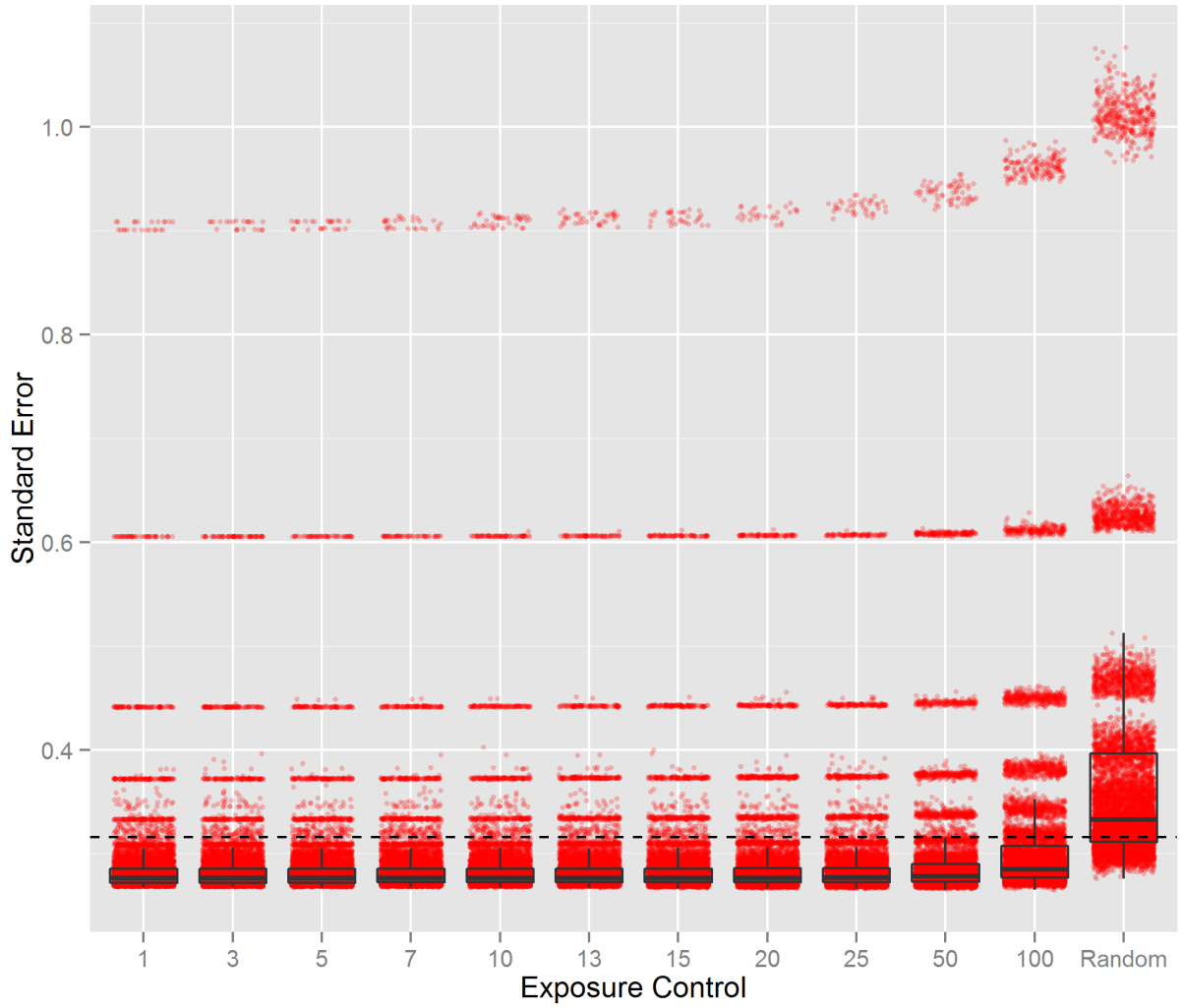

Figure 5.25: Standard Error Distribution by Exposure Control Condition

Exposure Rates Exposure rate distribution of each exposure control condition is shown in Figure 5.27. In this figure, one observation with exposure rate 1 in Rand-1 condition has been removed from the figure to show the spread better. Since there was not an exposure control in Rand-1 condition, the same first item has been selected as a first item for each simulee. For this reason the exposure rate of that item is 1. In the figure, bold lines in the middle of the box plots shows the median of exposure rates. The median exposure rate increased as the amount of randomization in item selection increased, except for Rand-100 condition. It is important to note that since the ratio of item pool size to the test length is constant

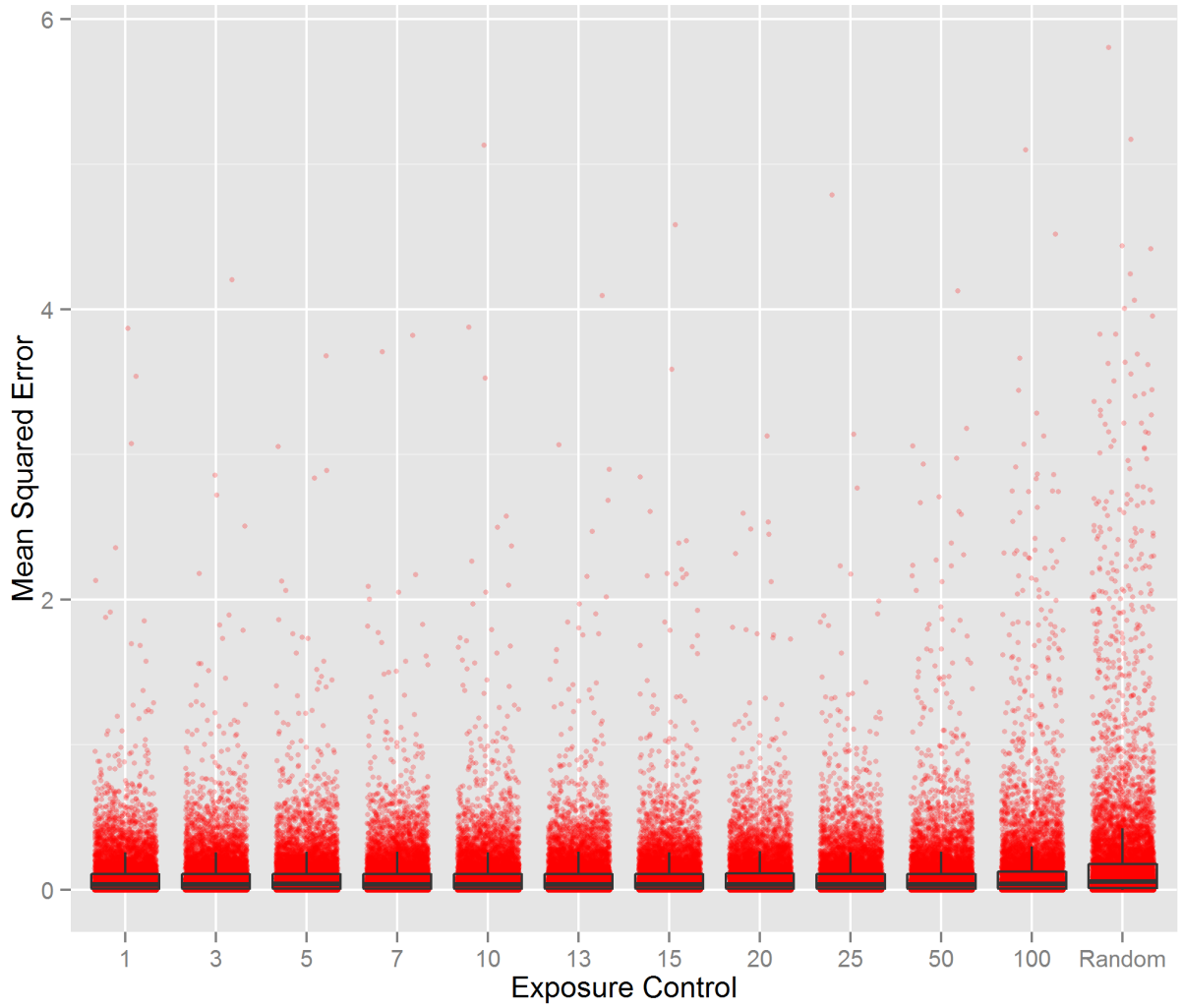

Figure 5.26: Mean Squared Error Distribution by Exposure Control Condition

for each condition, the mean exposure rates were equal for each condition (Table 5.6). It can be observed that exposure control method worked. The number of highly exposed items decreased as the amount of randomization imposed on item selection increased. For total randomization condition, all items exposed at the similar rate.

Table 5.6 shows the means and standard deviations of exposure rates in addition to the ratio of exposure rates that are larger than 0.20 and smaller than 0.05. Even though the means of exposure rates were equal, the standard deviations of exposure rates decreased steadily except for condition Rand-100. A small standard deviation of exposure rates means,

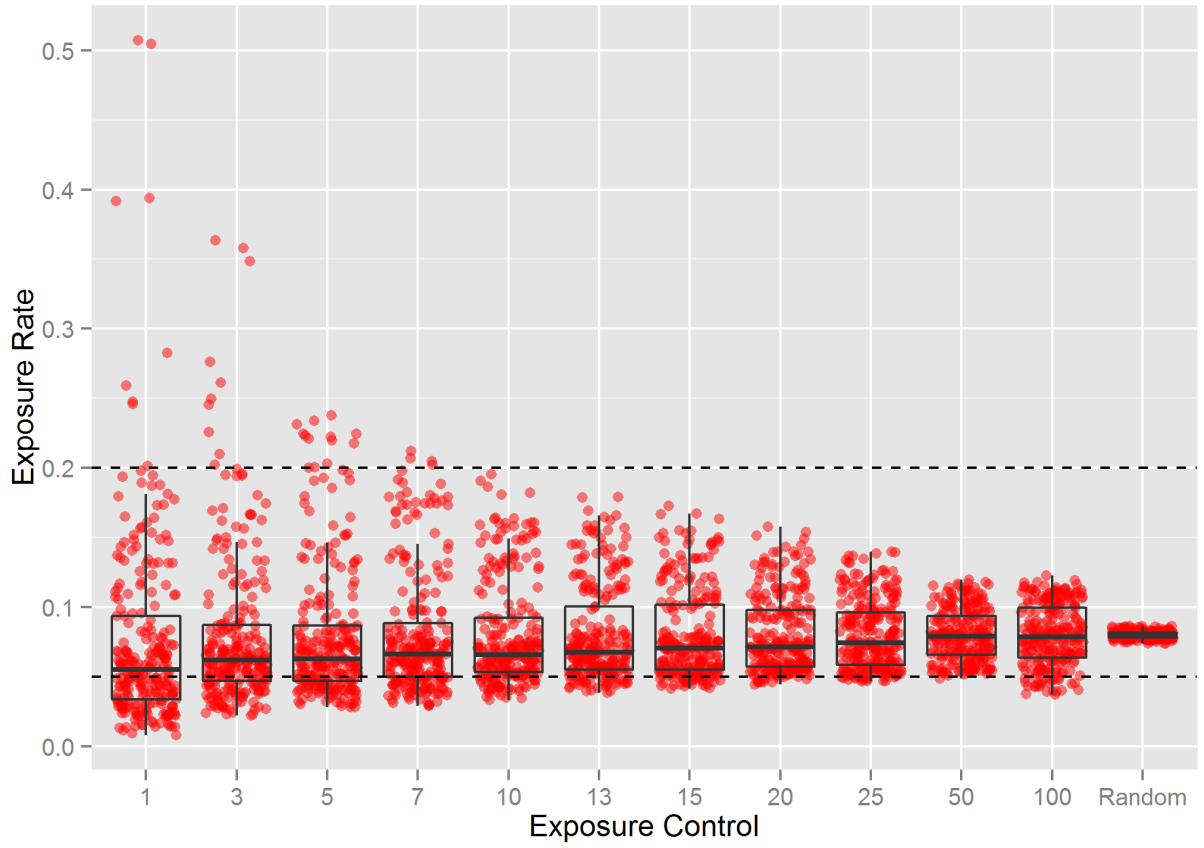

Figure 5.27: Item Exposure Distribution by Exposure Control Condition

the uniform exposure of items across examinees. The standard deviation of exposure rates is important because as Chen et al. (2003) showed in their paper (Equation 14, p. 134), there is a direct relationship between the variance of exposure rates and the average between-test overlap rate. In the CAT scenario here, where the mean exposure rates were equal between conditions, the reduction in the standard deviation of exposure rates directly translates to a reduction in the average between-test overlap rates. This means, for conditions where standard deviation of exposure rates were lower, the ratio of items two randomly selected examinees both see would be lower. This is an important test security issue for an operational high stakes CAT test.

The percentage of items that were exposed more than 20% of the examinees reduced as the amount of randomization in item selection increased. After the condition Rand-10, none

Table 5.6: Item Exposure Analysis by Exposure Control Condition

Exposure Control	Mean Exposure	SD	Exposure > .20	Exposure < .05
1	0.0800	0.0916	0.0400	0.4560
3	0.0800	0.0566	0.0400	0.2800
5	0.0800	0.0490	0.0480	0.2920
7	0.0800	0.0439	0.0160	0.2520
10	0.0800	0.0378	0.0000	0.1600
13	0.0800	0.0343	0.0000	0.1280
15	0.0800	0.0323	0.0000	0.1120
20	0.0800	0.0276	0.0000	0.0520
25	0.0800	0.0247	0.0000	0.0600
50	0.0800	0.0180	0.0000	0.0040
100	0.0800	0.0216	0.0000	0.1000
Random	0.0800	0.0025	0.0000	0.0000

of the items exposed more than 20% of the examinees. The percentage of items that were exposed to less than 5% of the examinees also reduced as the amount of randomization in the item selection increased, except for Rand-100 condition. This indicates that, all of the items were utilized to a greater extent and item pool became more efficient. These results showed that, from the exposure control perspective, total randomization of item selection did the best job. But as previous results regarding bias, SE and MSE suggested this condition is not desirable from other the aspects of the measurement practice.

IPUI Previous paragraphs showed that the mean values of bias, MSE and SE did not differ for exposure control conditions between Rand-1 to Rand-25. On the other hand, Figure 5.22 shows that IPUI decreased steadily as more randomization imposed on item selection mechanism. For conditions between Rand-50 and Rand-Total, the decrease is very clear. Table 5.5 shows this decrease numerically. The standard deviation of IPUI values increased steadily as well, except for Rand-Total condition.

Figure 5.28 displays the distribution of IPUI values for each exposure control condition visually. The increase in the spread of IPUI values are apparent from the box plots shown. The bold lines in the box plots shows the median values of IPUI. Median values were generally

larger than the average values given in Table 5.5. Consequently, the decrease in the median values cannot be seen from the graph.

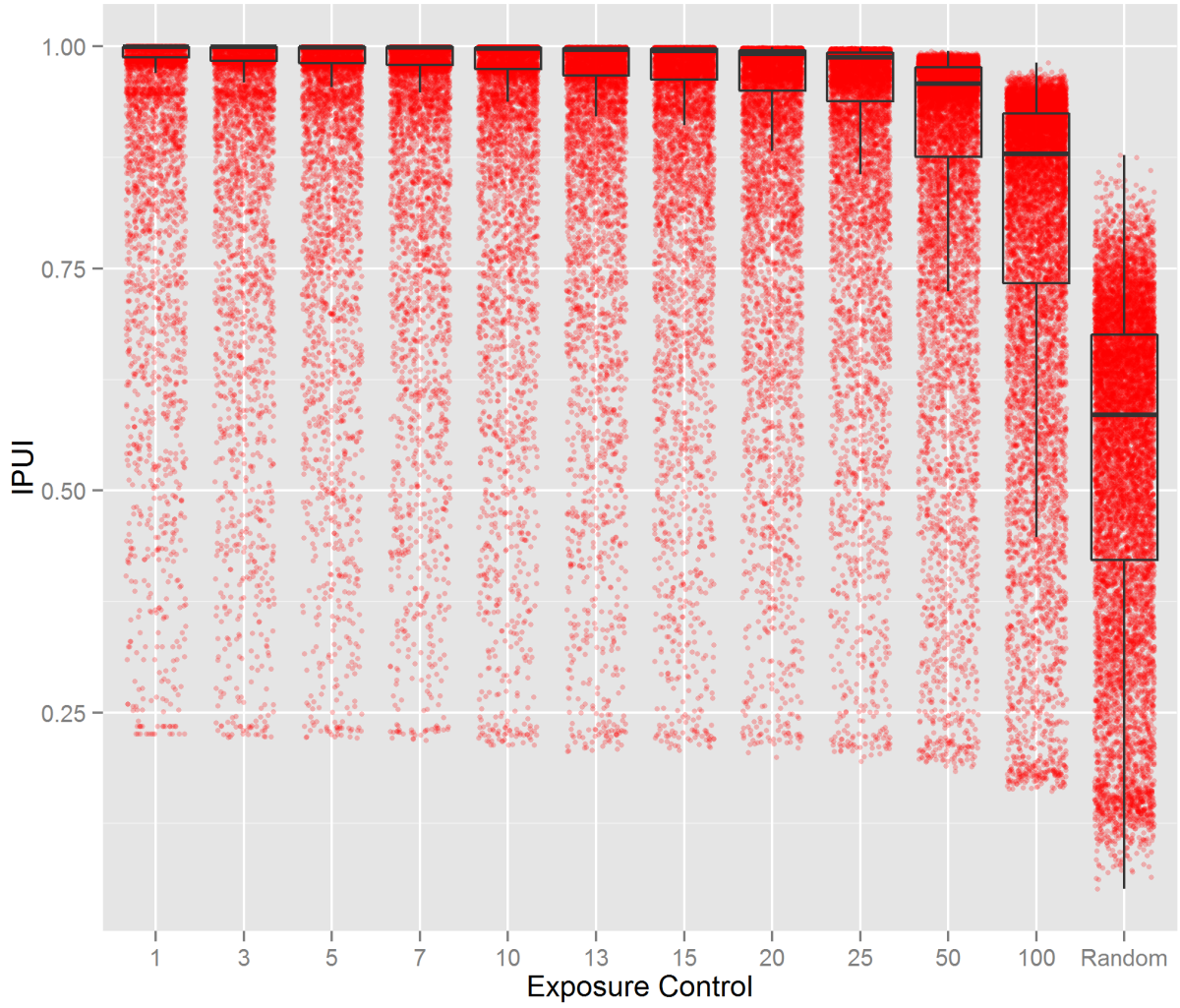

Figure 5.28: IPUI Distribution by Exposure Control Condition

For the Rand-Total condition, even though the overall IPUI values were smaller, majority of the examinees had IPUI values larger than 0.6. The main reason for this was the overlap of item pool and ability distribution. If there was a large discrepancy between item pool and ability distribution, then the overall IPUI values would be even less.

IPUI and Standard Error The relationship between SE and IPUI is shown in Figure 5.29. For each condition there was a negative relationship between SE and IPUI. The relationship was curvilinear and there were more spread in the values of IPUI compared to SE. The relationship between bias and IPUI is plotted in Figure D.3 on page 185. There was not a clear relationship between these two variables.

5.1.3 Research Question 3

In this research question, the use of IPUI as a diagnostic tool is evaluated. A hypothetical example of a state testing agency was introduced in Section 4.2.1.4 on page 44. This state testing agency had three plans to test. Item pool information of these item pools were shown in Table 4.1 on page 46, the distributions of item pools are in Figure 5.30. The first plan consist of an item pool of size 90 from three different content areas. The overall difficulties of each content area were different. In this first plan, content balancing was imposed on the item selection. In the second plan, the same item pool used for Plan 1 was used. Different than Plan 1, content balancing was not imposed on the item selection. In Plan 3, the item pool consisted of 90 items from the difficult content area. No content balancing was imposed on the item selection for this plan as well. Other than the difference between item pools and the content balancing, the CAT specifications of the plans were the same.

The following paragraphs investigate these item pools using traditional CAT outcomes such as mean bias and mean SE at each true θ. These results are compared to the IPUI results and the utility of IPUI as a diagnostic tool is discussed. Since this research question investigates the item pools, in the following pages, plans are referred as item pools. So, Plan 1, 2 and 3 are referred as IP-1, IP-2 and IP-3, respectively.

Bias Figure 5.31 shows the mean bias values conditional on true θ values for each item pool. IP-1 and IP-2 performed similarly throughout the ability scale. IP-1 had overall higher bias compared to IP-2. This shows the effect of content balancing. IP-3 had similar mean

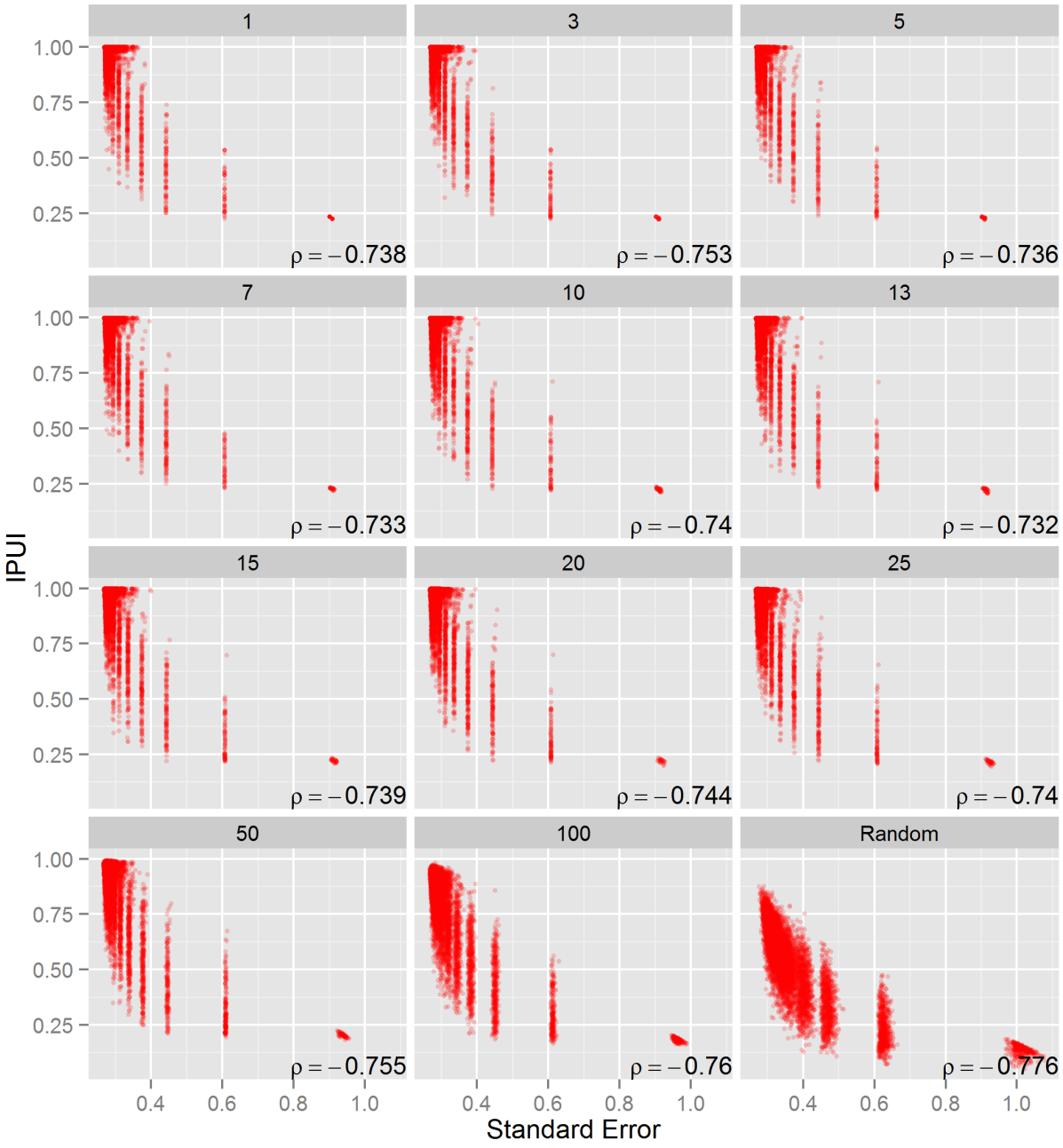

Figure 5.29: IPUI and Standard Error Relationship by Exposure Control Condition

bias on the positive side of the ability scale. On the negative side of the ability scale, the mean biases were much larger compared to the other item pools. As Figure 5.30 shows, IP-3 did not have any items that had difficulty parameters less than 0.5. Clearly, this reflected on the biases of ability estimates. The bias distributions for each item pool condition shown in

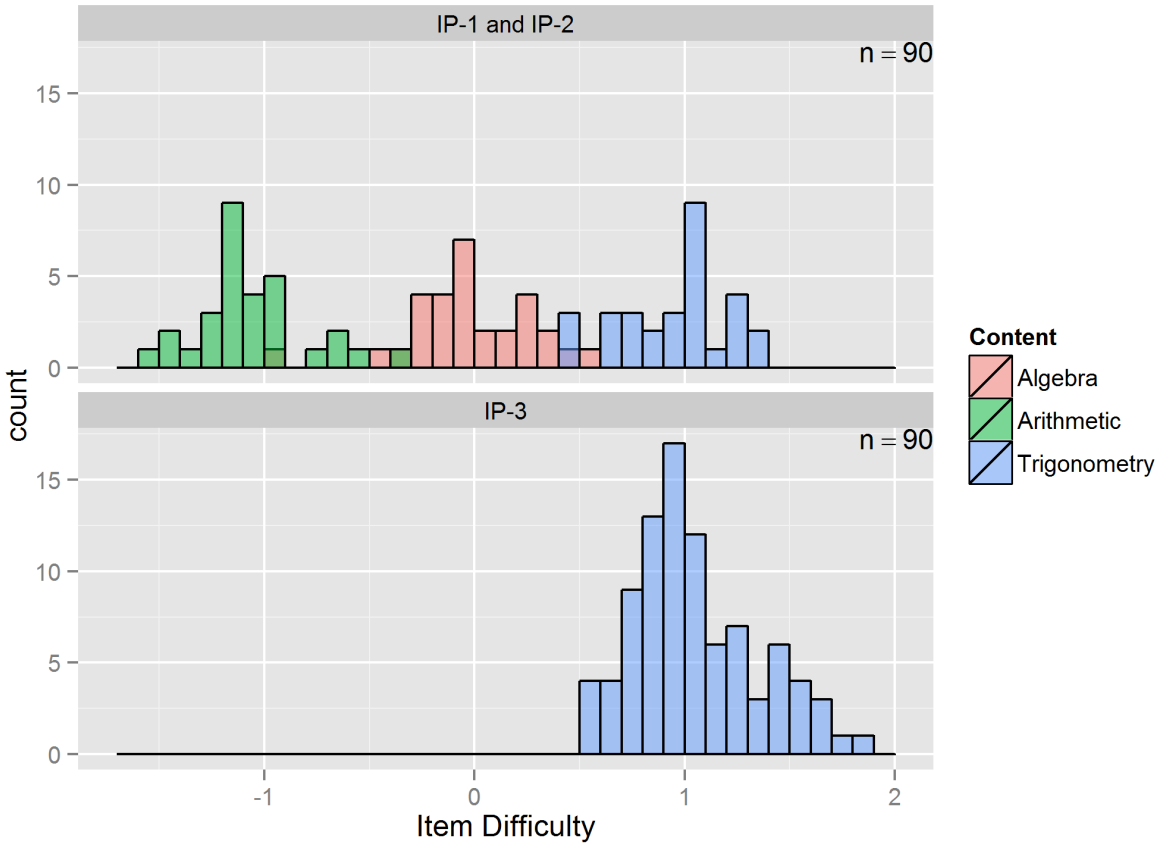

Figure 5.30: Item Pool Distributions of Proposed Test Plans

Figure E.1 on page 186 corroborates this observation.

Standard Error Figure 5.32 shows the mean standard error values conditional on true θ values for each item pool. IP-2 in plan 2 performed the best throughout the ability scale except the positive end of the ability scale. The mean SE values for this item pool were the lowest at this interval. The mean of SE for IP-1 of plan 1 had the same shape as the IP-2 but had larger values compared to the IP2. Close to the middle of the ability scale, the difference between these two item pools almost disappeared. Mean SE values of IP-3 were substantially large at the negative side of the ability scale. At the right of $\theta = .5$, where this item pool was comparatively strong, it had the lowest SEs. Figure E.2 on page 187 shows the distributions of the SEs at each true θ value. This graph also corroborates the mentioned

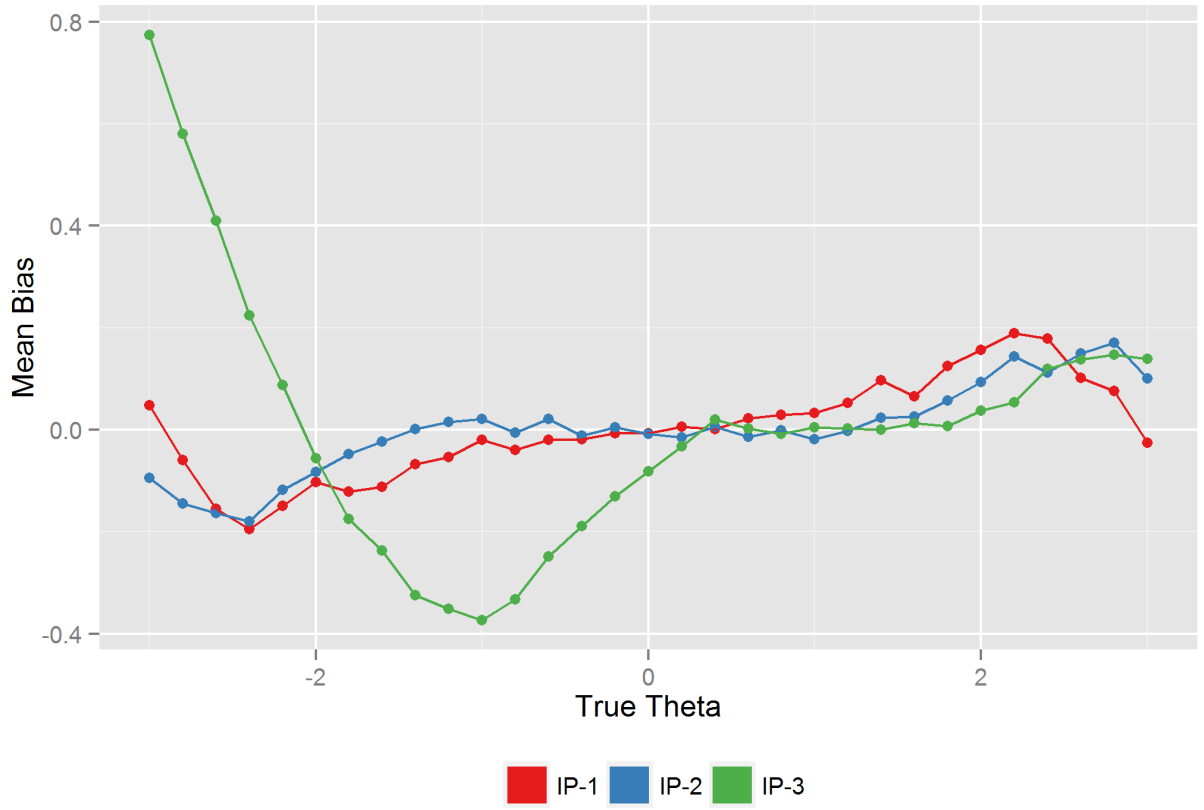

Figure 5.31: Mean Bias Conditional on True θ for each Plan (Item Pool)

differences among the item pools.

IPUI Figure 5.33 shows the mean IPUI values conditional on true θ values for each item pool. Performance of the IP-2 of plan 2 was better than other item pools for θ values smaller than 1. For true θ values larger than 1, the performance of IP-3 was better. The mean IPUI values of IP-1 was lower than the IP-2 for all true θ values. This clearly shows the effect of content balancing on the performances of the item pools.

The similarity of SE and IPUI can be observed from the figures. If the lines in Figure 5.33 were reflected along the x axis, the shapes would resemble the SE distributions on Figure 5.32. But the differences between these two figures are also apparent. For example, even though IPUI results show a large difference between the performances of item pools for Plan 1 and

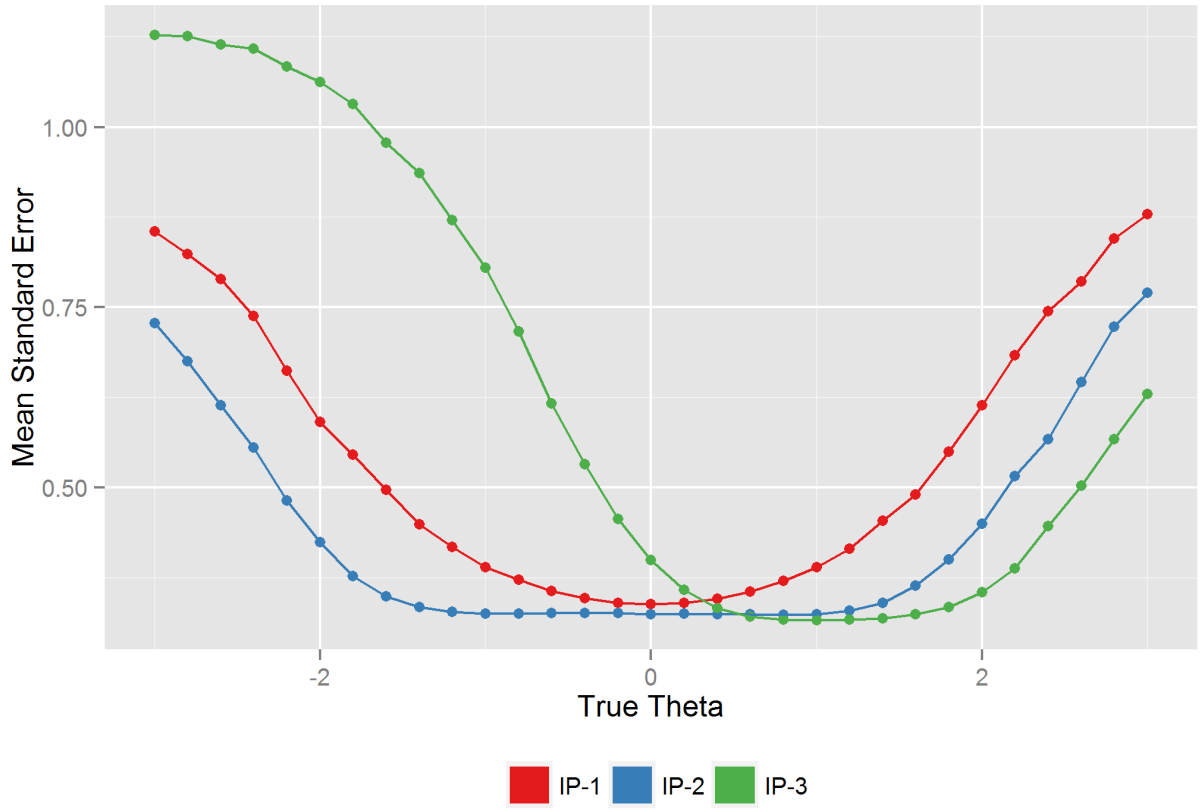

Figure 5.32: Mean Standard Error Conditional on True θ for each Plan (Item Pool)

Plan 2, this difference was not reflected on the SE for θ values between -0.5 and 0.5. So, for the true θ value 0, the absolute difference between IPUI values of IP-1 and IP-2 were about 0.2, the absolute difference of mean SE values were almost non-existent, 0.014. On the other hand, for $\theta = 2.2$, the absolute differences between both the mean IPUI values and the mean SE values were 0.17. So, there is not a one to one relationship between these two indicators.

Figure 5.33 shows the weak points of the item pools for each plan. It is advisable for test developers to add more items around the θ values where IPUI values were low. But when there is content balancing as in Plan 1, the advise of adding more items to the item pool where IPUI values are low might not be clear. Because, a test developer might ask "Add items from which content area?". To answer this question, the test developer might check the graph of the mean IPUI value at each test question for a given true θ.

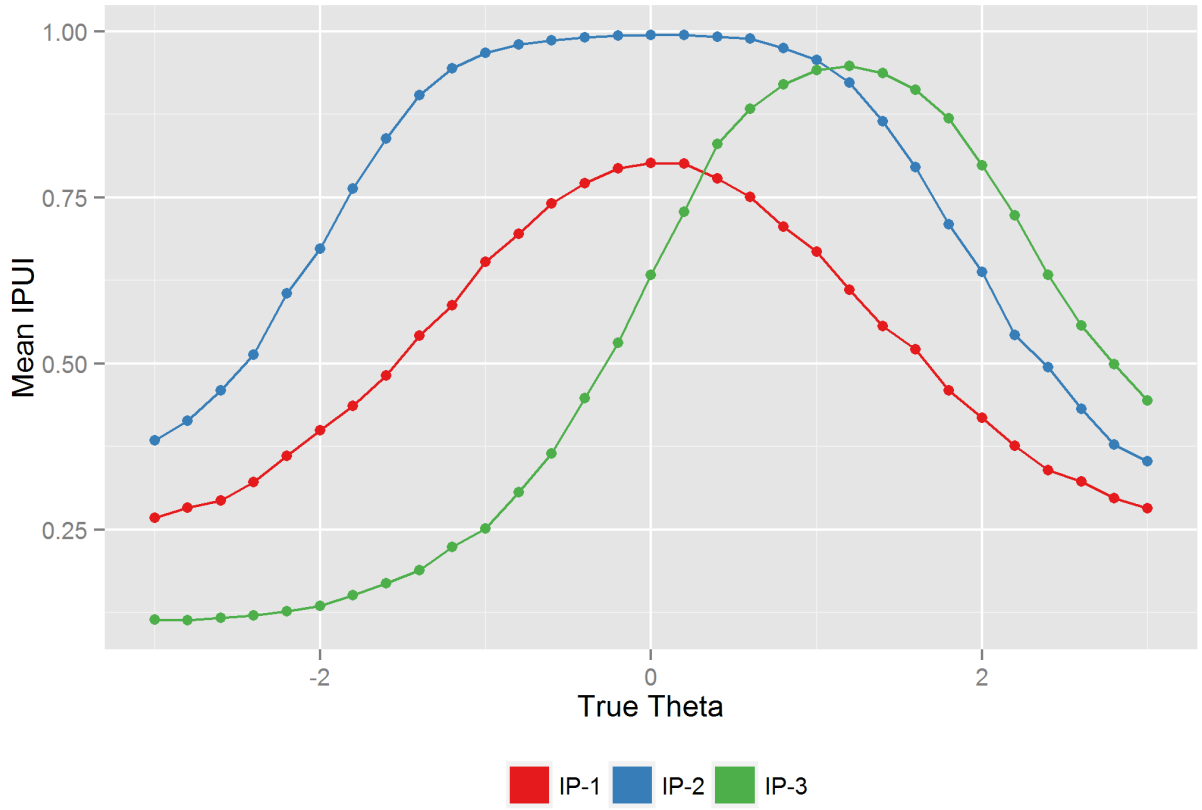

Figure 5.33: Mean IPUI Conditional on True θ for each Plan (Item Pool)

Figure 5.35 shows the graph of the mean IPUI at each item number of the test. In this graph a test developer can observe whether the item pool's performance reduces as the test proceeds for certain groups of examinees. If the item pool cannot provide appropriate items, the IPUI value will decrease as the test proceeds. Such information will guide the test developers about how many items are needed to add to the item pool to resolve this deficiency. For example, IP-2 able to provide appropriate items to the examinees with true θ values between -1.4 and 1.4 throughout the test. But at the extremes of the ability scale, after the second item, the item pool failed to provide appropriate items. At around 13 items, the items did not match the ability estimate of the examinees with extreme true θ values.

IP-2 and IP-3 shows a smooth declining trend from the beginning of the test to the end for all of the true θ values. But IP-1 shows a zigzag shape. The reason behind this is the

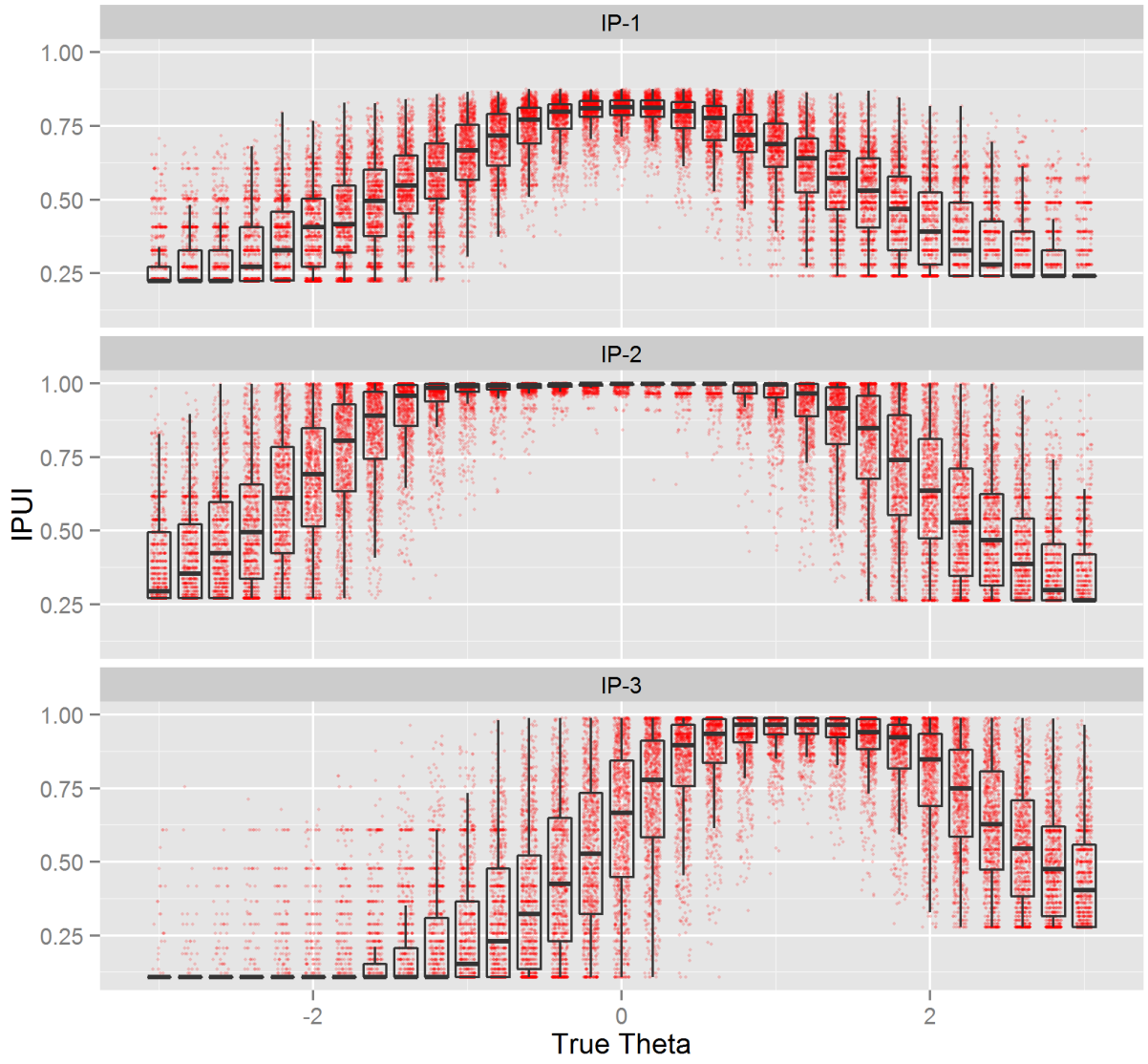

Figure 5.34: IPUI Distribution Conditional on True θ for each Plan (Item Pool)

content balancing. For example, for true θ value 1.4, the CAT algorithm on average gave an item with IPUI value of 0.24 as the fourth item in Plan 1. But the next item had a higher IPUI value, the sixth item had even higher IPUI value. For item 4, even though there were many appropriate items available in the item pool, there was not a difficult item in content area 1 (arithmetic) to present to the examinees. Content balancing combined with a weak item pool distorted the motivation behind the CAT: providing the most appropriate items to the examinees. The problem of content balancing when there is an interaction between item

difficulty and content areas was also described in Way et al. (2002):

> If some item types are inherently more difficult than others and the content specifications call for every examinee to be administered equal numbers of each, the algorithm will tend to choose the most difficult items from the easy content areas for the high-ability examinees. (p. 146)

A test developer can look at the Figure 5.35 and add items to the content areas which has low IPUI values. For item pools without content balancing (IP-2 and IP-3), this graph shows a similar information as the previous two graphs. But still this graphs gives an idea to the test developer about approximately how many items should be added to the item pool around the vicinity of the θ to make it satisfactory.

Figure 5.36 shows the relationship between the intermediate θ estimates of each examinee and the IPUI value of the particular item administered that is appropriate for that intermediate θ estimate. The points are color coded based on the content of the item administered. This graph combines all of the examinees at each true θ condition. In this regard, the ability distribution of the examinees can be seen as a uniform distribution. Since there was no content balancing in other plans, the color coding is relevant only for plan 1.

This graph is very helpful for test developers to see the weak spots of the item pool. For plan 1, due to the content balancing imposed, IPUI values were low when the item selection algorithm selected an item from a content area where the intermediate θ estimates were outside the difficulty range of that content area. For instance, when an examinee's intermediate θ estimate was low and item selection algorithm had to administer a trigonometry item, the IPUI value for that intermetiate θ estimate would be low, even though the easiest trigonometry item available administered to the examinee. As test proceeds the item pool depleted easy trigonometry items and even harder trigonometry items administered to the examinee. This vicious cycle continues until the test ends. A test developer can observe this from Figure 5.36 and add easy trigonometry items to the item pool. Algebra items

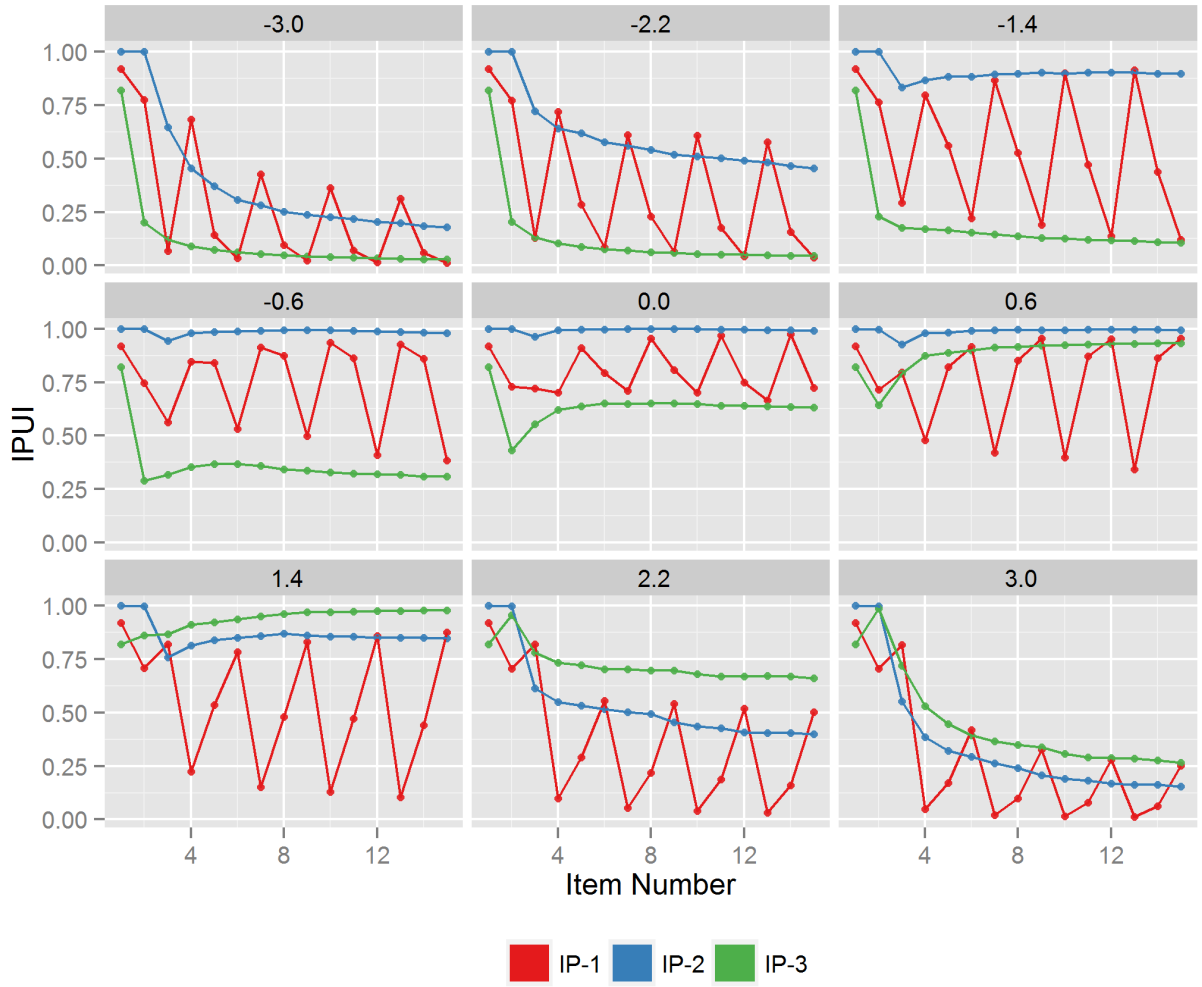

Figure 5.35: Mean IPUI at each Item Number for Selected True θs

were adequate around the middle of the θ scale but more algebra items needed outside the θ interval [-1, 1].

The same item pool performed better in plan 2 when there was no content balancing imposed. The item pool was adequate when examinee's intermediate ability estimates were between -1.5 and 1.5. This item pool needs very easy and very difficult items.

The item pool for plan 3 performed well between 0.5 and 2. Outside this interval, this item pool failed to provide appropriate items to the examinees. More items are needed for this item pool with item difficulty parameters equal to the intermediate θ estimates where

IPUI values were low.

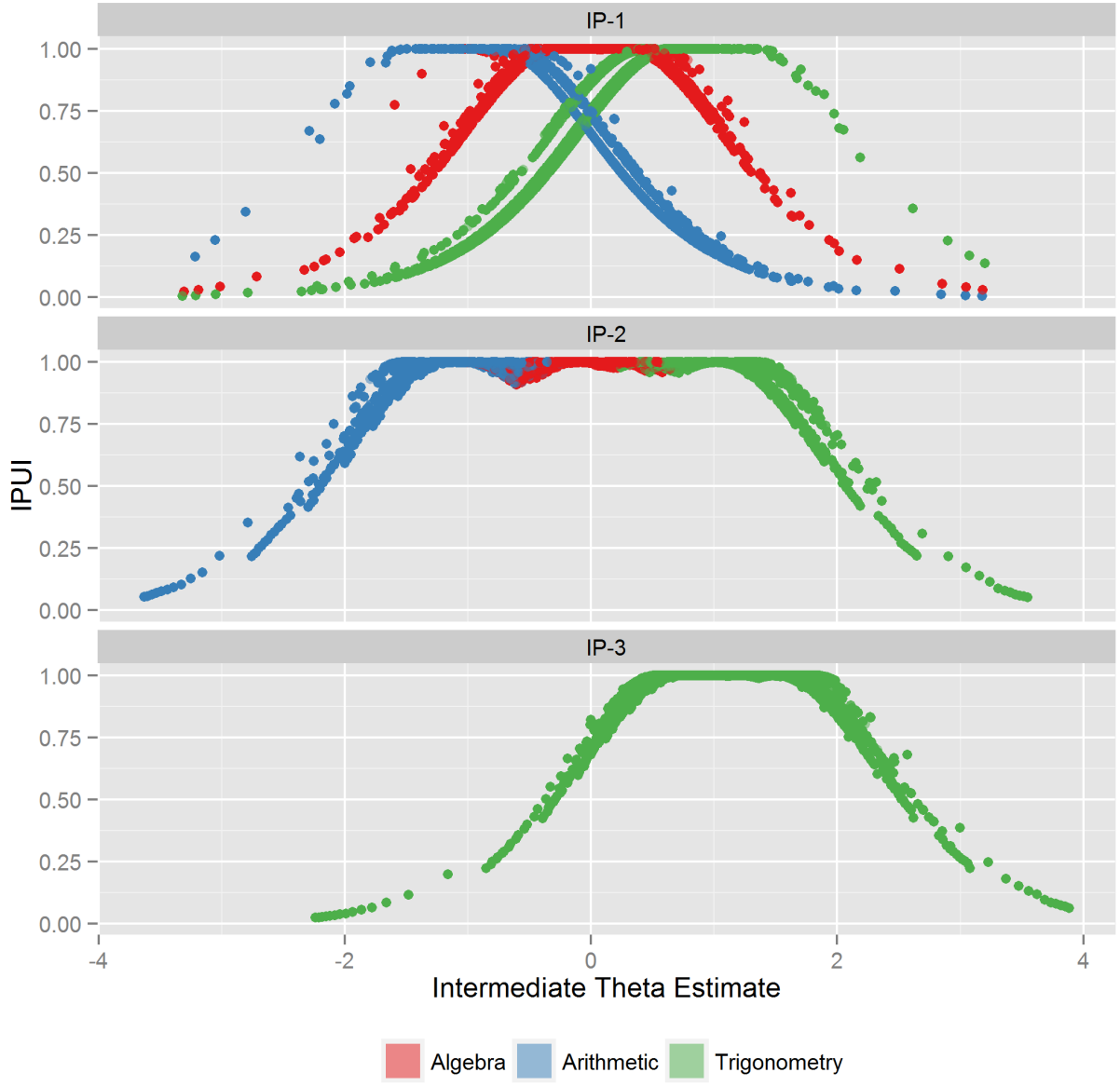

Figure 5.36: The Relationship between Intermediate θ Estimate and IPUI

One limitation of this graph is it's dependence on the quality of the θ estimates. For plan 3, even though there were many examinees that had true θ values between -3 and -2, almost none of these examinees had θ estimates that were lower than -2. In fact, as Figure 5.31 on page 99 shows, there was a large positive bias in the estimates for plan 3. This might cause a

test developer to think that no items are needed with item difficulty parameter values smaller than -2. Adding easier items to this item pool would reduce the biases in the θ estimates and potentially highlight the need for much easier items.

Figure 5.37 shows the relationship between intermediate θ estimates and the difficulty parameters of the items administered at those θ estimates. The points are colored to indicate the level of the IPUI values. The green points represent high IPUI values and red points represent low IPUI values. The dashed red line is the identity ($y = x$) line. If the item pool provided appropriate items to the examinees, the points would fall on the identity line.

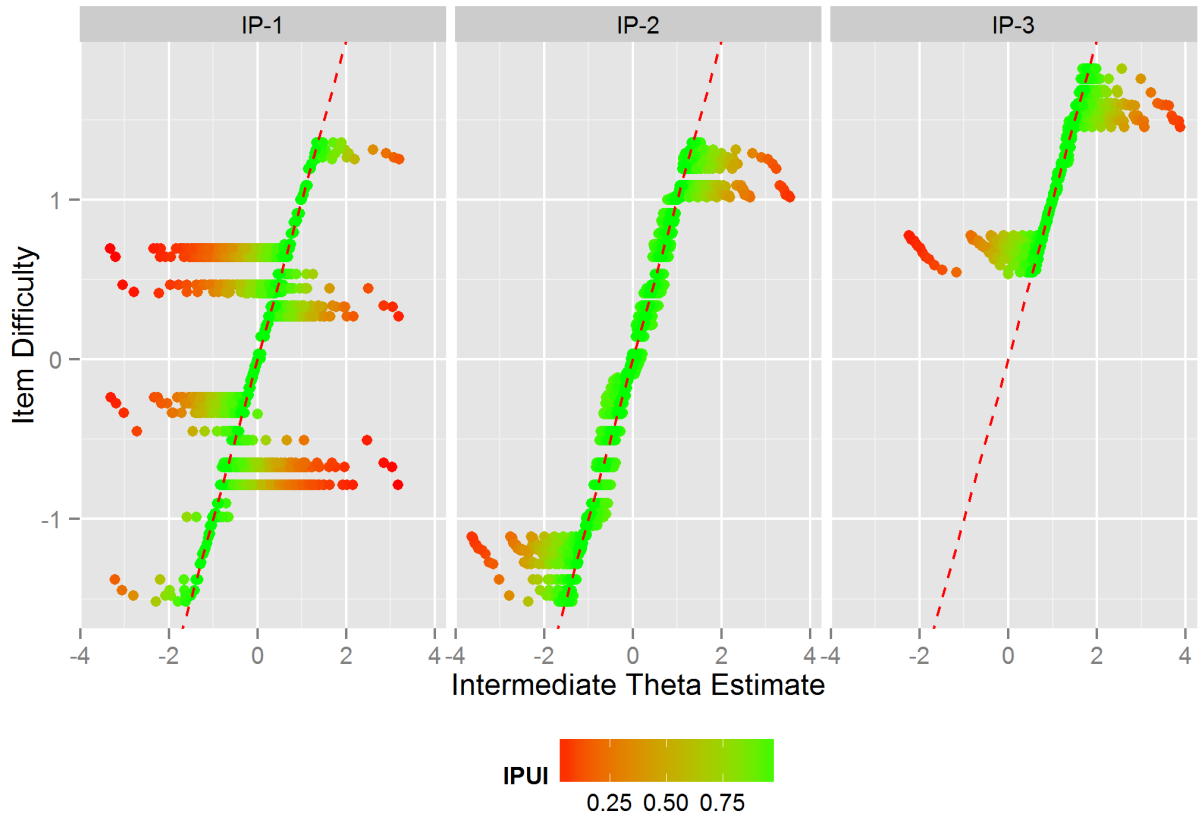

Figure 5.37: The Relationship between Intermediate θ Estimate and Item Difficulty

The graph for plan 1 shows the effects of content balancing. The points deviated from the identity line throughout the θ scale. This item pool was adequate between -2 and 3 for plan 2. The deviations started towards the extremes. This item pool was not adequate when

the intermediate θ estimates reached the extremes of the θ scale. The item pool for plan 3 was adequate for only a small portion of the θ scale. This item pool was not adequate for the θ estimates outside the interval between 0 and 2.

5.2 Second Phase - Operational Data

In the second phase of the analysis, Research Questions 4 and 5 were investigated. These two research questions were answered using five operational item pools in addition to four generated item pools. Two of the generated item pools were ideal item pools generated for the specifications and examinee population of NCLEX-RN examination. Other two item pools were generated from the first operational item pool by removing some proportion of the items. In the following sections, first the results of the item pool generation results will be presented, other two sections will present the results of the Research Questions 4 and 5.

5.2.1 Ideal Item Pool Generation

Two ideal item pools were generated to compared with the operational item pools. The first ideal item pool had fixed bin sizes with lengths 0.4. The middle bin was centered around 0 in the θ scale. The second ideal item pool had fixed bin sizes with lengths 0.8. The first item pool was designed to be more precise compared to the latter. Reducing the length of the bin sizes further would increased the precision but in that case more items would be needed. If the item pool had more than necessary items, most of them would not be administered to the examinees. This would reduce the efficiency of the item pool. Previous research (He & Reckase, 2013) and previous unpublished item pool design studies for NCSBN showed that fixed bin sizes with 0.4 and 0.8 bin widths would be good choices.

In the creation of the ideal item pools, 10,000 examinees were simulated. Each simulee needed different items depending on her θ value and responses. The size of the ideal item pool grew as the number of simulees increased. After a number of simulees, there was a lot

of overlap between the items needed by simulees and the need for new items decreased unless a simulee had an extreme ability or irregular responses. Figure 5.38 shows the growth of the size of the ideal item pool for fixed bin size 0.4. The growth graph for fixed bin size 0.8 is in Figure F.1 on page 188. The first panel of the graphs show the growth progress for each content area. Second panel of the graphs show the overall growth progress of the ideal item pool.

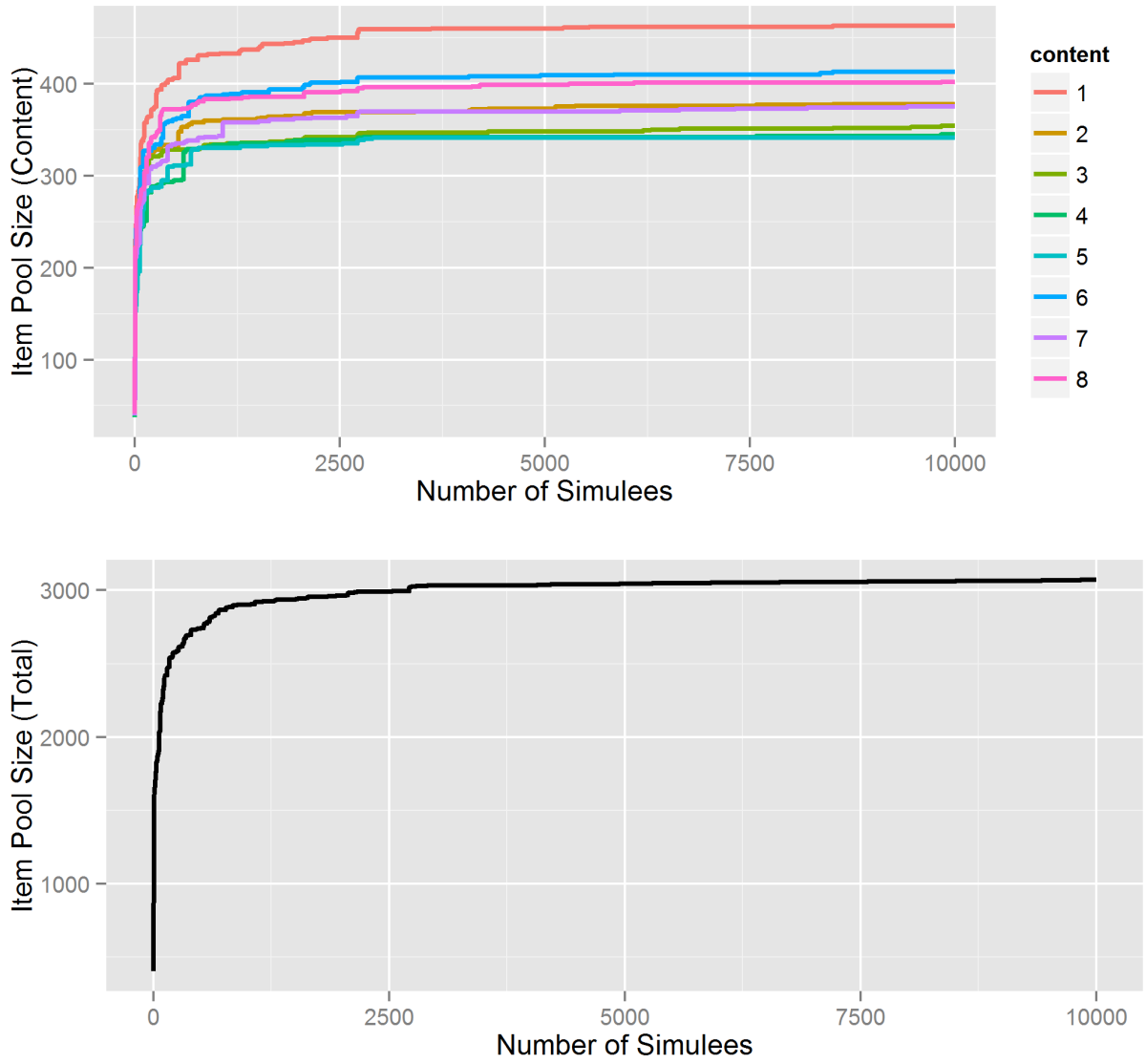

Figure 5.38: Progress Plot for Ideal Item Pool with Fixed Bin Size 0.4

These growth progress graphs are important in the diagnosis of the bin-and-union method.

If the lines do not converge to a single point, this means the number of examinees simulated was not enough for the creation of the ideal item pool. In that case, more simulees are needed to obtain a stable ideal item pool. For the ideal item pools generated for this dissertation, the convergence reached after 3000 simulees for the fixed bin size 0.4 item pool and after 2000 simulees for the fixed bin size 0.8 item pool.

At the end of the simulations, ideal item pool with fixed bin size 0.4 had 3071 items. Ideal item pool with fixed bin size 0.8 had 1652 items. The item difficulty parameter distribution by content area for the ideal item pool with fixed bin size 0.4 is in Figure 5.39. The same graph for the fixed bin size 0.8 is in Figure F.2 on page 189. The b-parameter distributions had almost the same shape for each content area. The number of the items within each content area reflects the content area distributions shown in Table 4.2 on page 47. The peaks towards the extremes of the ability scale reflect the relatively large number of items within the last bins. The bins at the extremes captured all of the items that were needed in the θ scale from the last bin's boundary to the infinity.

5.2.2 Item Pools Used in the Second Phase

In the second phase of this study, five operational item pools (named as Op 1,..., Op 5) were compared to four generated item pools. Two of the generated item pools were ideal item pools, the ideal item pool with bin size 0.4 (Ideal 0.4) and the ideal item pool with bin size 0.8 (Ideal 0.8). The last two item pools were generated by removing half of the first operational item pool (Half-IP) and by removing two thirds of the first operational item pool (One-Third IP).

The item difficulty distributions and the number of items within each item pool is in Figure 5.40. The dashed red line in the histograms are the mean item difficulty values of each item pool. The number of items within each operational item pool was the same. The distributions of the operational items were almost the same as well. The majority of the items were close to the cut score. There were more easy items than difficult items within

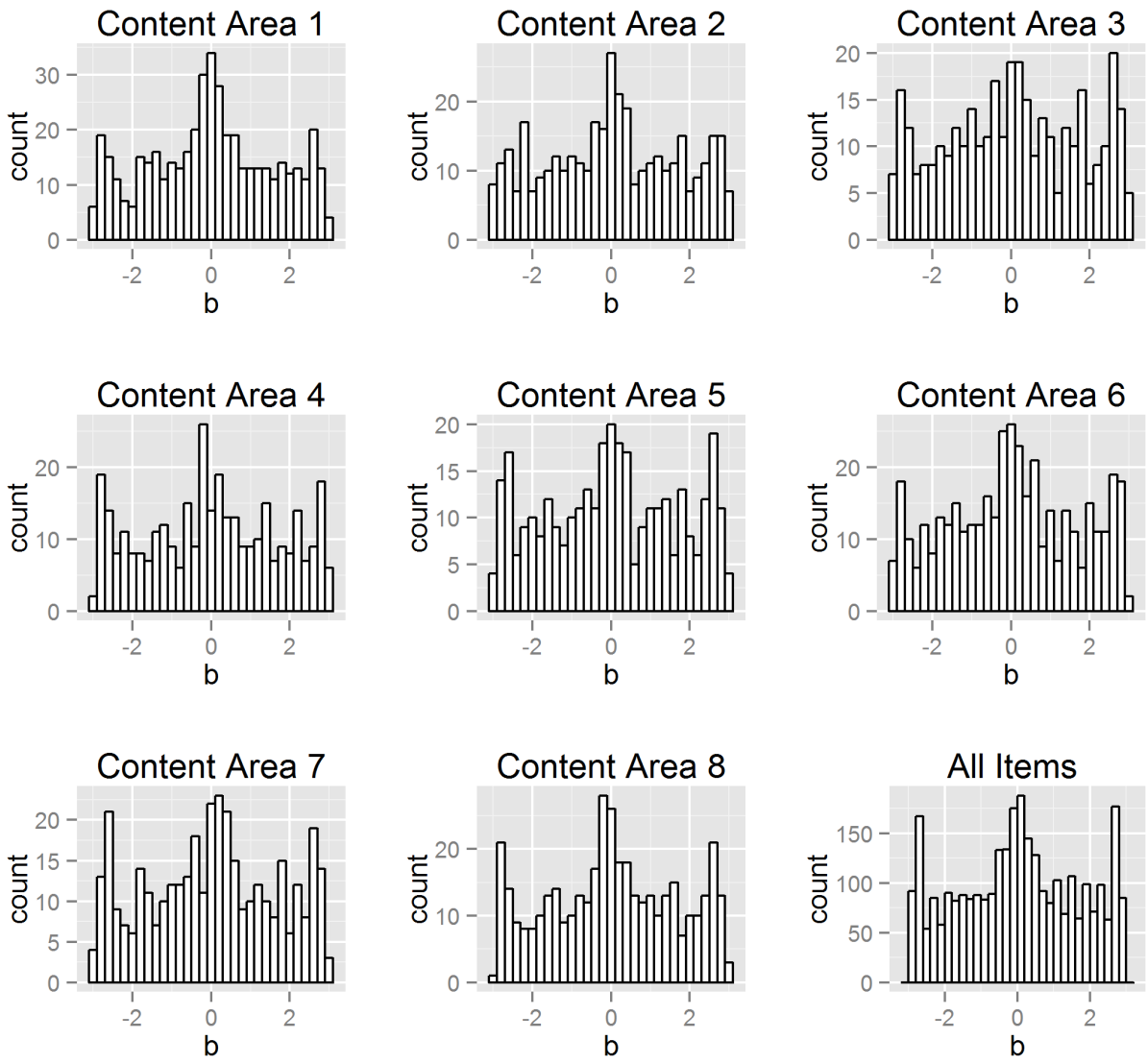

Figure 5.39: Item Difficulty Distributions by Content Area for Ideal Item Pool with Fixed Bin Size 0.4

each operational item pool. Also the operational item pools did not spread much to either extreme of the ability scale.

The half item pool and the one third item pool had exactly half and one third as many items as the first operational item pool, respectively. The shape of the distributions of these two item pools were similar to the operational item pools.

The ideal item pool with fixed bin size 0.4 had more than twice as many items as the

operational item pools. This item pool had many items close to the middle of the ability scale. Different than the shapes of the operational item pools, this item pool had almost equal number of easy and hard items. The spread of this item pool was also wider compared to the operational item pools.

Ideal item pool with fixed bin size 0.8 had 180 items more than the operational item pools. It had less items around the middle of the ability scale compared to both operational item pools and Ideal 0.4 item pool. It had more items at the extremes. The boundaries of the bins are visible for this item pool. Close to the bin boundaries, there was a bump in the number of items. While generating the ideal item pools, the candidate items were generated from a standard normal distribution. If the items were generated from a uniform distribution, the bumps close to the bin boundaries would have been disappeared. But if a uniform distribution was used as the generating function, the decision for the minimum and maximum value of this distribution would have been arbitrary. The parameters of the uniform distribution cannot be infinity, as a result the ability scale would have been confined arbitrarily.

5.2.3 Research Question 4

The aim of this research question is to observe the performance of IPUI for an operational CAT. In the previous research questions, the CAT scenarios were rather simplistic. Conditions differed in only one aspect from a very simple CAT algorithm. In this research question, a complex CAT algorithm was investigated. This CAT algorithm includes content balancing, exposure control, a two stage ability estimation method and a complicated stopping rule. Predicting the item pool performance for rather simple CAT scenarios might be feasible. But for a complex CAT algorithm as the one investigated in this research question, predicting item performance would be difficult.

The same simulee sample was used for the simulations for each item pool condition. 10,000 simulees were generated from a normal distribution with the similar means and standard

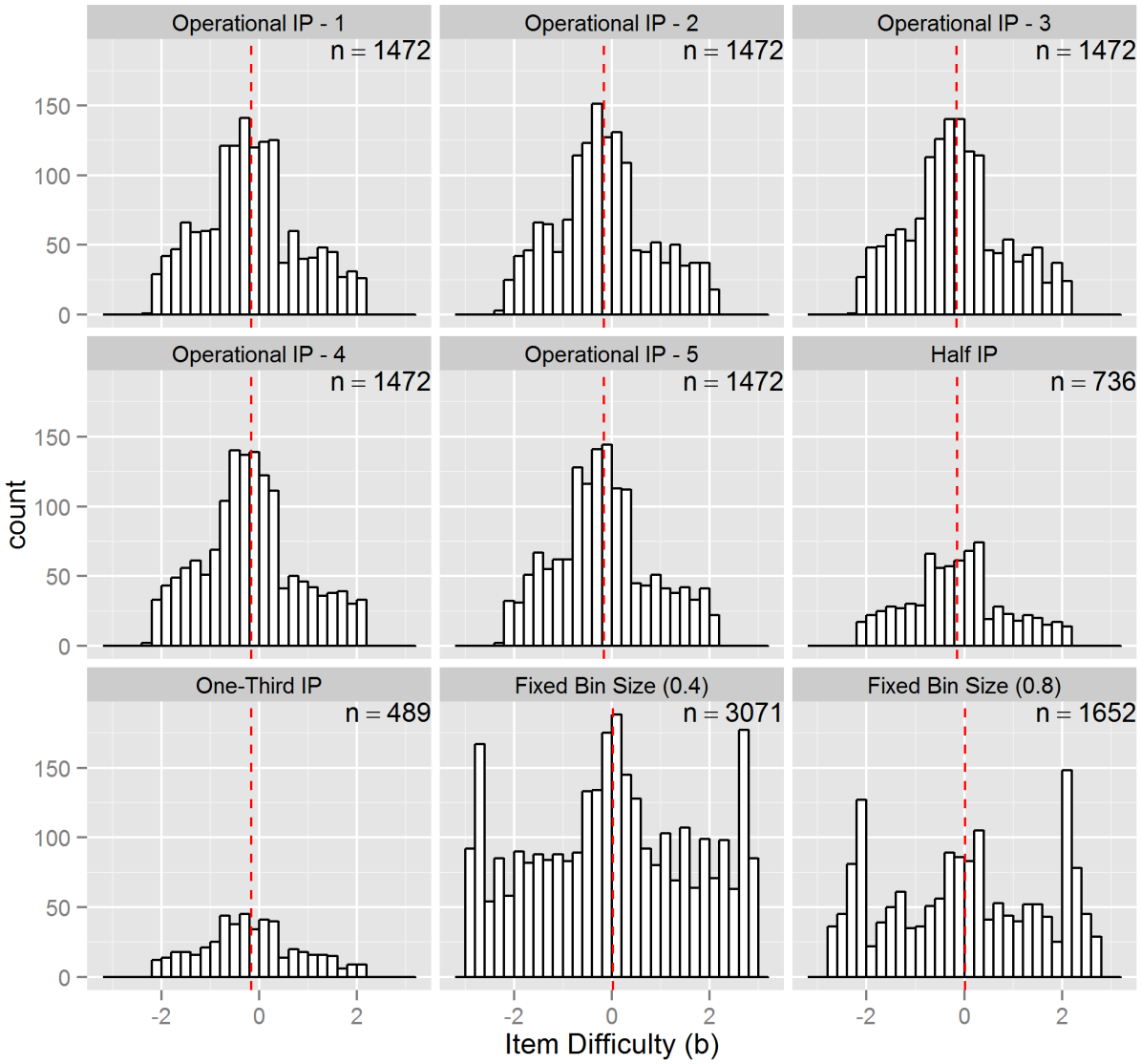

Figure 5.40: Item Difficulty Distributions for Item Pools Used in Research Question 4

deviations of the real examinees that took the tests. The θ distribution of the simulees are plotted in Figure G.1 on page 190.

For each item pool, means of bias, SE, MSE, IPUI, decision accuracy and fidelity coefficient calculated. These values for each item pool condition are given in Figure 5.41. As can be seen from the graph, except for the IPUI values there was almost no difference between different item pools. Table 5.7 shows the numeric values of means and standard deviations of these values. Table 5.7 also shows the mean and standard deviation of the test length for each

item pool condition. In the following pages, each of these outcome variables will be discussed separately.

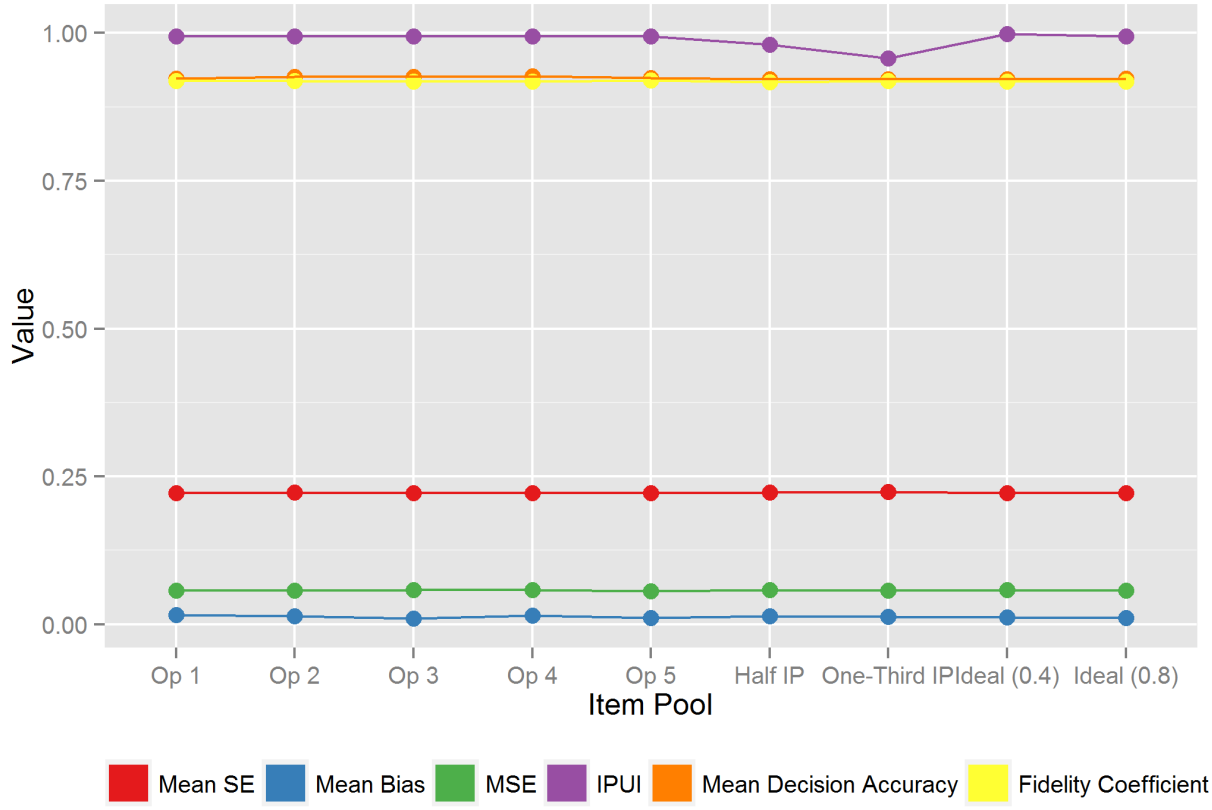

Figure 5.41: Summary Statistics for Research Question 4

IPUI The size of the item pool did not make much difference in other outcomes of the adaptive test except the exposure rates (which will be discussed later in this section). Since the IPUI is a direct measure of the quality of an item pool, the differences in the item pools are expected to reflect on the IPUI values. In Figure 5.41, the only difference between item pools was their IPUI values. The IPUI values of Half-IP and One-Third-IP were lower compared to the other item pools.

Table 5.7 shows no difference between the mean and standard deviations of the IPUI values for the operational item pools. Ideal-0.8 item pool had a similar mean IPUI value but

Table 5.7: Summary Statistics for Research Question 4

Item Pool	Fidelity Coefficient	Bias	SE
Op 1	0.9184	0.016 (0.238)	0.222 (0.055)
Op 2	0.9192	0.014 (0.239)	0.223 (0.055)
Op 3	0.918	0.010 (0.241)	0.222 (0.055)
Op 4	0.9178	0.015 (0.240)	0.222 (0.055)
Op 5	0.9197	0.011 (0.236)	0.222 (0.056)
Half IP	0.917	0.014 (0.241)	0.223 (0.056)
One-Third IP	0.9191	0.012 (0.238)	0.224 (0.056)
Ideal (0.4)	0.918	0.012 (0.240)	0.222 (0.055)
Ideal (0.8)	0.9181	0.011 (0.239)	0.222 (0.056)

Item Pool	MSE	IPUI	Test Length	Decision Accuracy
Op 1	0.057 (0.092)	0.995 (0.007)	108.556 (72.913)	92.34%
Op 2	0.057 (0.093)	0.995 (0.006)	108.524 (73.285)	92.59%
Op 3	0.058 (0.093)	0.995 (0.007)	108.629 (72.866)	92.61%
Op 4	0.058 (0.091)	0.995 (0.007)	109.007 (73.083)	92.68%
Op 5	0.056 (0.092)	0.995 (0.007)	109.635 (73.565)	92.43%
Half IP	0.058 (0.093)	0.980 (0.019)	109.446 (73.228)	92.21%
One-Third IP	0.057 (0.091)	0.957 (0.030)	110.220 (74.077)	92.23%
Ideal (0.4)	0.058 (0.095)	0.999 (0.001)	108.718 (72.903)	92.22%
Ideal (0.8)	0.057 (0.091)	0.994 (0.002)	109.352 (73.183)	92.33%

Note. Numbers within the parentheses are standard deviations of each outcome.
SE: Standard Error; MSE: Mean Squared Error

the standard deviation of IPUI values was lower compared to the operational item pools. Ideal-0.4 item pool had a mean IPUI value that is almost 1 and the standard deviation of IPUI values was the lowest. Half and One-Third item pool conditions had lower mean IPUI values and they had larger standard deviations. The quality of the item pools were clearly reflected on the statistics of the IPUI values.

Figure 5.42 shows the distributions of IPUI values visually. Operational item pools had similar distributions but they had many examinees who had comparatively lower IPUI values. Ideal-0.8 item pool had far fewer examinees that had lower IPUI values. None of the examinees got an IPUI value less than 0.95 for this item pool condition. Ideal-0.4 item pool performed the best. All examinees, except 5 of them, had IPUI values larger than 0.99. Spread of IPUI values were much larger for the Half-IP and One-Third-IP conditions. In

fact, none of the examinees in One-Third-IP condition had an IPUI value larger than 0.99.

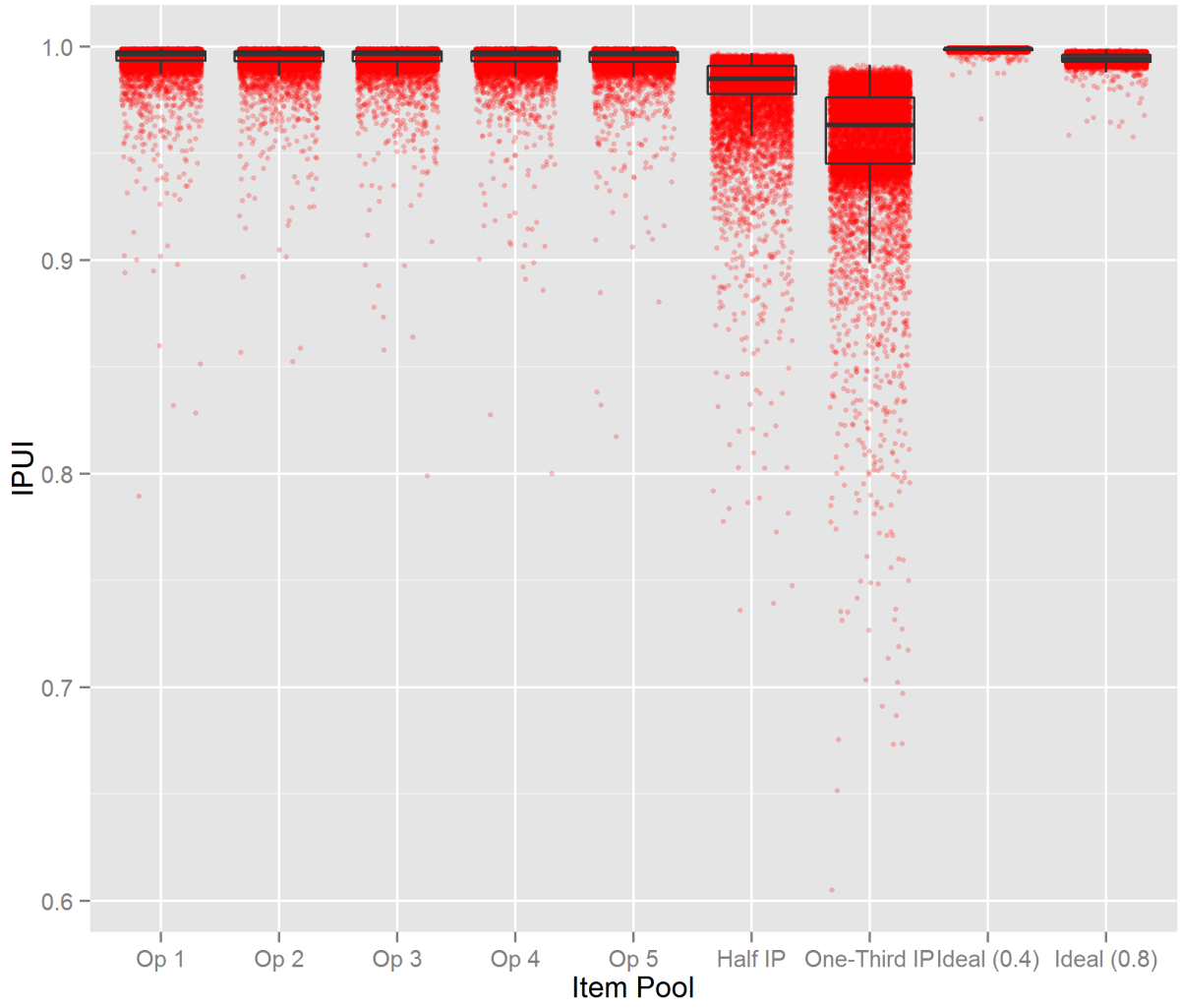

Figure 5.42: IPUI Distribution for each Item Pool Condition

Relationship between True and Estimated Ability Figure 5.43 shows the relationship between true ability (θ) and estimated ability ($\hat{\theta}$). The dashed lines in the figure are the identity lines (i.e. $y = x$ line). At each condition, there was an error in the estimation and points deviated somewhat from the identity line. Close to the middle of the ability scale there is a squeeze in the spread around the identity line, estimated abilities approximated their true values better. The reason of this is the variable length test. Around the cut score,

test lengths were longer compared to the examinees at the extremes of the ability scale. As indicated at Section 5.1.2.1 on page 74, longer tests increase the test precision and reduce the estimation error. Figure 5.43 also shows the correlations between true and estimated abilities for each item pool. There was almost no difference between the correlations of the item pools.

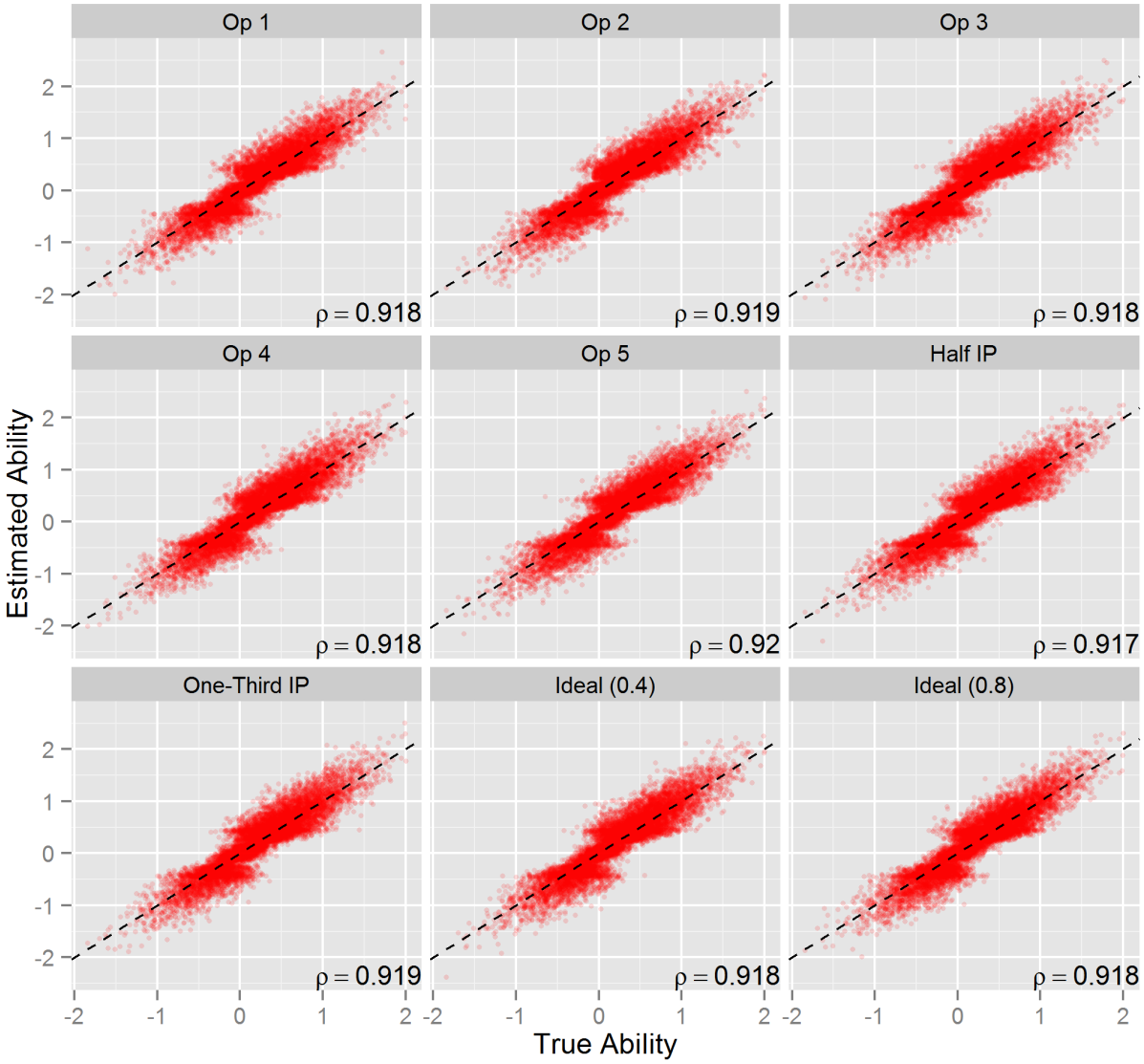

Figure 5.43: The Relationship between True θ and Estimated θ

IPUI and Estimated Ability Figure 5.44 shows the relationship of IPUI and estimated ability. This graph is useful to see the locations in ability scale where the item pool was

insufficient. Operational item pools had very high IPUI values around the cut score. But towards the extremes of the ability scale the spread of IPUI values increased. Especially for high ability examinees, the item pool did not have enough appropriate items.

Ideal item pools had high IPUI values throughout the ability scale. Half-IP had lower IPUI values for examinees close to the cut score. Towards the extremes of the ability scale the IPUI values decreased even more, especially for high ability examinees. One-Third-IP performed even worse than Half-IP. Close to the cut score, IPUI values make a dip, then towards the extremes of the ability scale IPUI values fell even further. Compared to the low ability examinees, high ability examinees had lower IPUI values.

Figure 5.44 showed that, except ideal item pools, the IPUI values of the examinees at the extremes of the ability scale were not as high as the examinees close to the cut score. From the perspective of test developers for this test, this might not be very problematic. The test is a licensure test with one cut score. As long as the decision accuracy is high for the test, the test is deemed to fulfill it's purpose. For this reason, the item pool should have sufficient items around the cut score. Test length can go up to 250 for examinees close to the cut score. The item pool should support this long test. Figure 5.44 gives some evidence for this. Around the cut score, for each item pool except the Half-IP and One-Third-IP conditions, examinees had almost perfect IPUI values.

IPUI and Test Length Additional evidence for the sufficiency of item pool around cut score comes from Figure 5.45 that shows the relationship between IPUI and test length. Except for Ideal-0.4 item pool, there was a spread of IPUI values for tests that last 60 items. Examinees with test lengths 60 were away from the cut score. Decisions for these examinees were clear after 60 items. These examinees were the ones that were located outside of the (-0.5, 0.5) band of the ability scale in Figure 5.44.

Tests were longer for examinees whose estimated abilities were close to the cut score. IPUI values for examinees who took tests longer than 60 items were close to 1 for operational

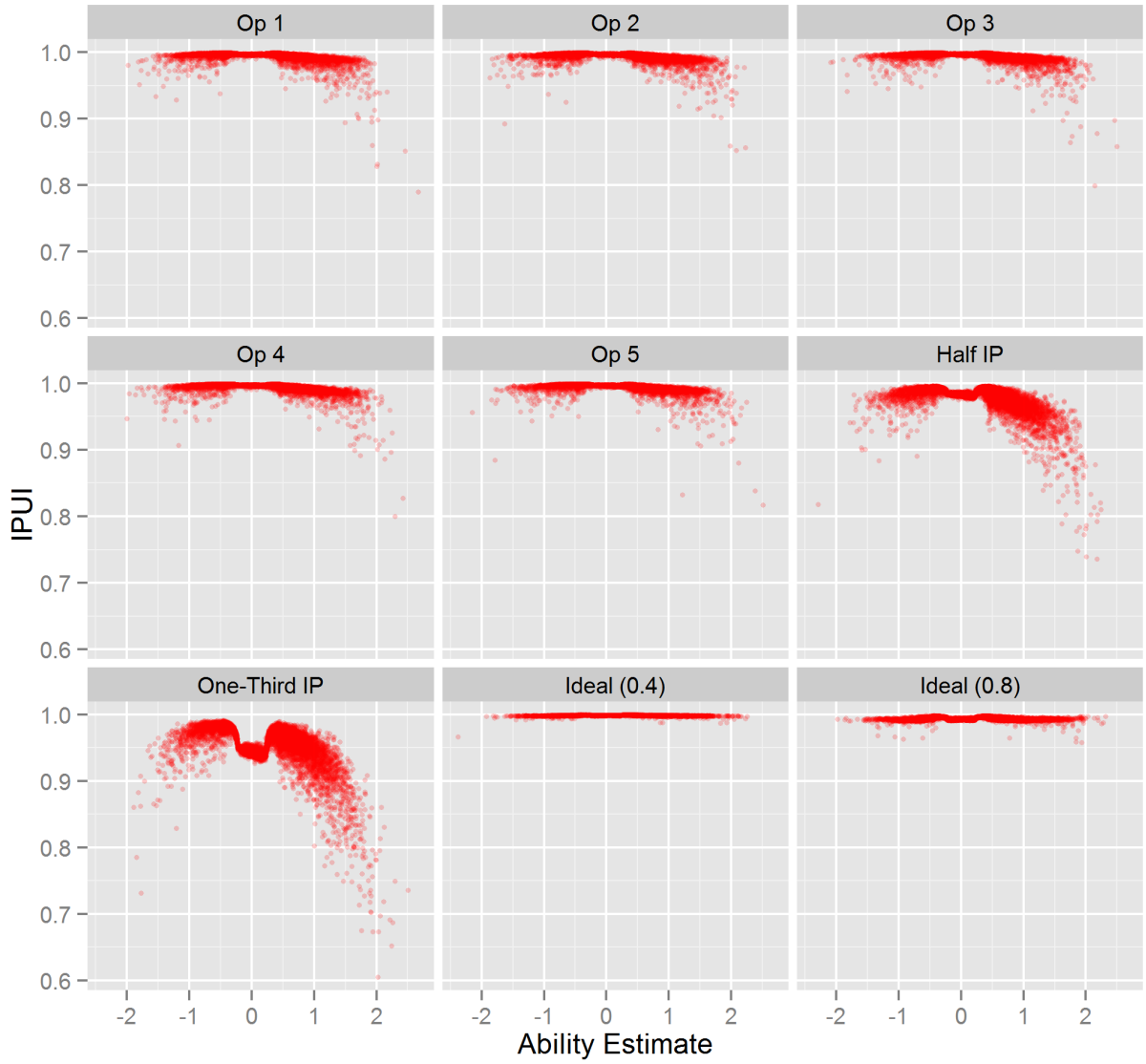

Figure 5.44: The Relationship between IPUI and Estimated Ability for each Item Pool Condition

item pools and ideal item pools. This is an important indicator for the quality of the item pool. These item pools were able to support very long tests with 250 items. It is crucial for an item pool to support long tests because high decision accuracy is much needed for the examinees who take these long tests. These tests were long because it was difficult to make a decision for these examinees. On the other hand, for Half-IP and One-Third-IP conditions, Figure 5.45 shows that IPUI values started to decrease as test length increased. This is more

visible for One-Third-IP. For these item pools, as test length increased, item pool failed to provide appropriate items.

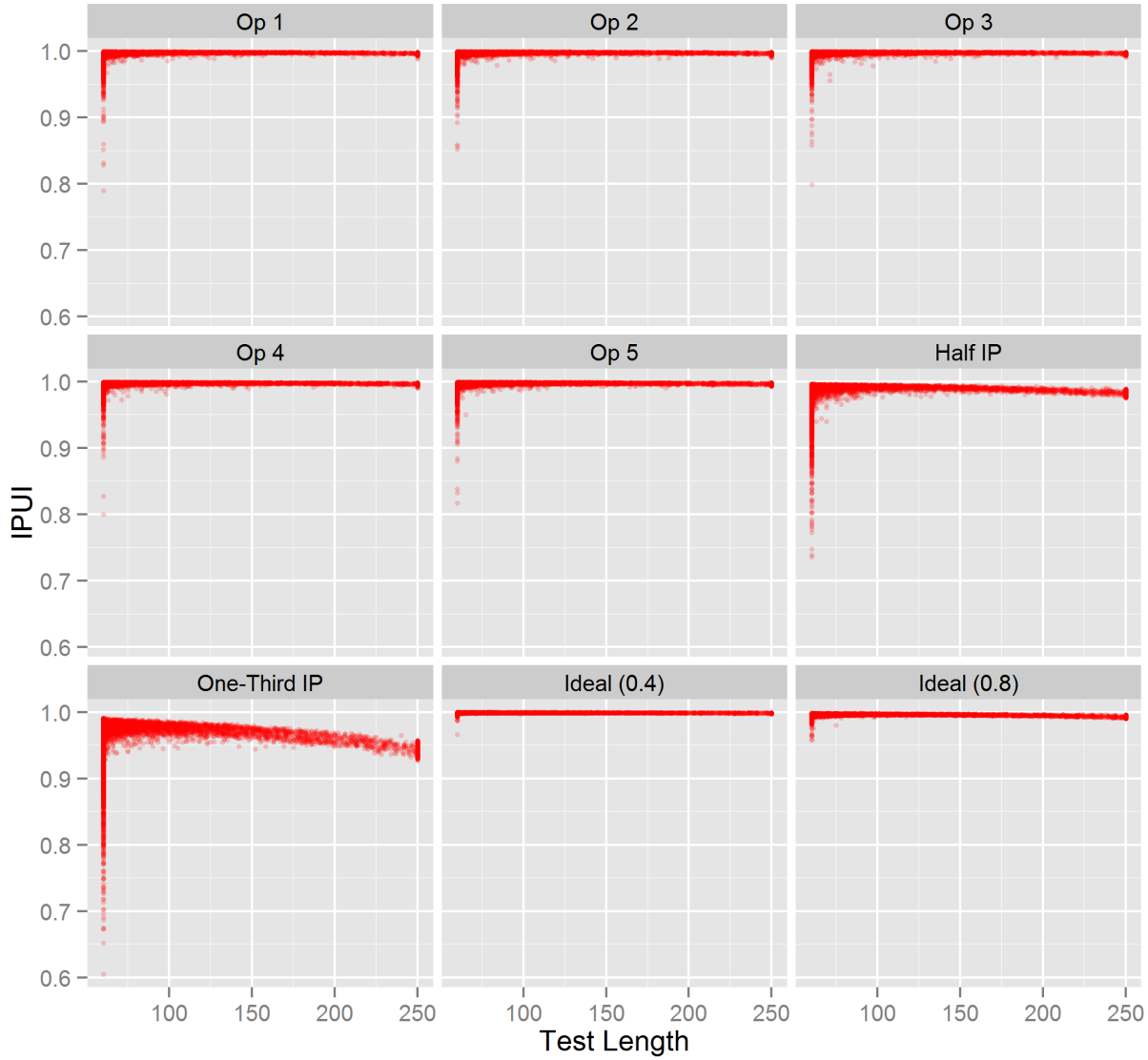

Figure 5.45: The Relationship between IPUI and Test Length for each Item Pool Condition

Bias Figure 5.41 shows no difference between the mean biases of the item pools. Table 5.7 shows that mean bias values were a little above 0 for each item pool condition. The standard deviation of the biases were the same for different item pool conditions as well. Figure 5.46

shows this visually. There was almost no difference between the bias distributions of different item pools.

Figure G.2 on page 191 shows the relationship between the estimated ability and bias. If there was no bias, the points would be on the dashed line in the middle. The bias was small around the cut score, as explained above. But, there was a rather large correlation between estimated abilities and biases for each item pool condition as shown in the text boxes. The regression lines fitted to the points also show this positive relationship. For high ability levels the biases were positive, abilities were overestimated. For low ability simulees, abilities were underestimated.

Bias and IPUI Figure 5.47 shows the relationship between bias and IPUI. A linear regression line was fitted to each plot to show the linear relationship between these two variables. From the graph it cannot be concluded that high absolute bias is associated with high IPUI. The expected relationship for this graph was an inverted-U shaped curvilinear line, where IPUI was high for low absolute bias values and low for high absolute bias values. This conclusion is in line with the findings of the previous research questions. On the other hand, the fitted lines shows that there is a weak negative relationship between bias and IPUI. For ideal item pools, this relationship almost disappears. This relationship is more evident for Half-IP and One-Third-IP conditions. Figure 5.44 shows that both operational and shortened item pools were weak on the positive side of the ability scale. Figure 5.47 reflects the weakness of the item pools for high ability examinees. When item pool distributions were balanced, as the ideal item pools, there was no relationship between IPUI and bias.

Standard Error Figure 5.41 shows almost no difference between the mean values of SEs for different item pool conditions. Numeric values at Table 5.7 also confirms this. Both the means and standard deviations of SE values were the same for each item pool condition. Only for the One-Third-IP condition did these values increase, but this increase was very small.

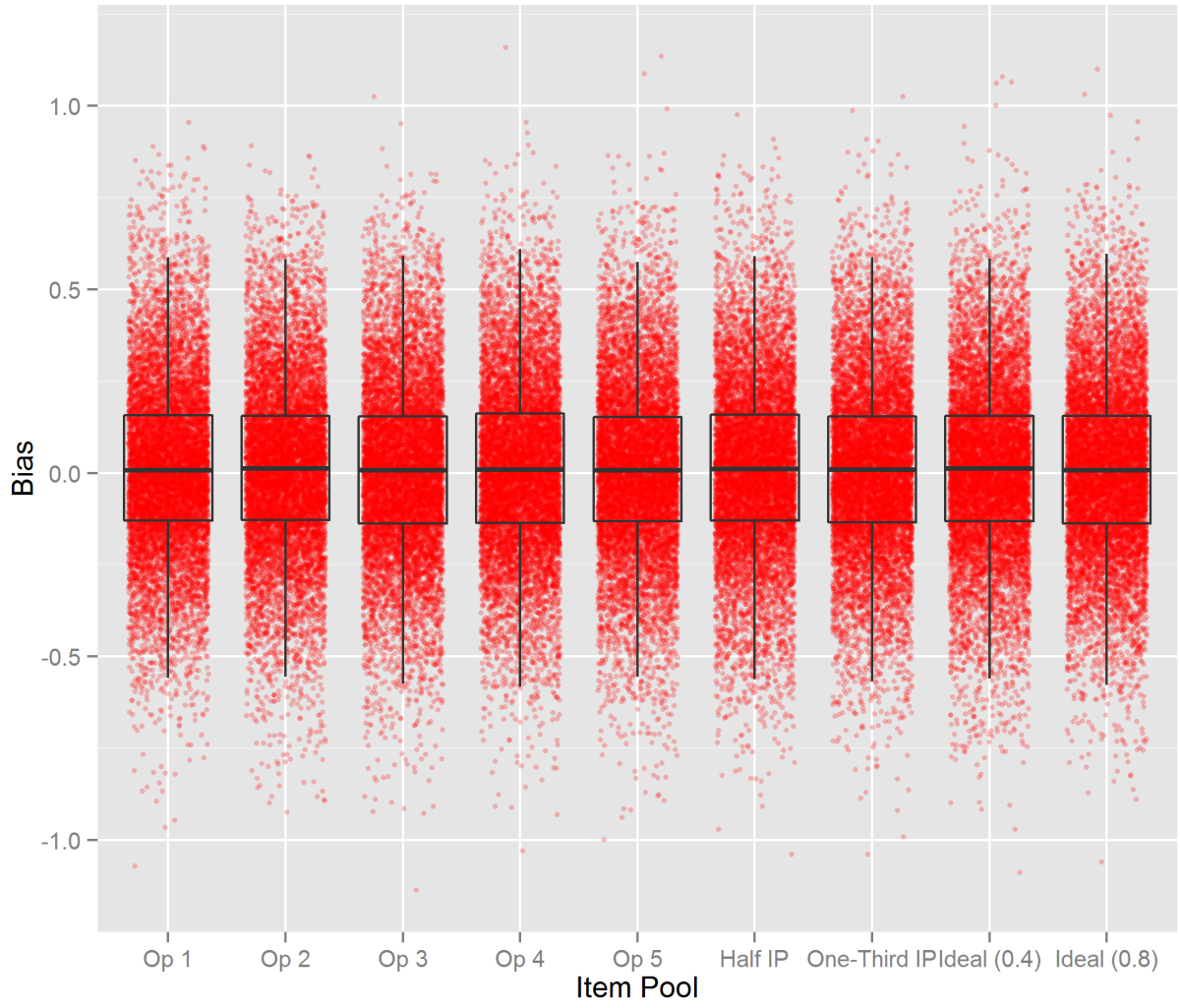

Figure 5.46: Bias Distribution for each Item Pool Condition

Figure 5.48 shows the distribution of SE values visually. The dashed line shows the threshold for a test with 0.9 reliability[1]. In this figure, except One-Third-IP, all of the remaining item pools had almost same SE distributions. The number of simulees who had high SE values were higher for One-Third item pool condition. There was one examinee who had a very large SE compared to the rest of the examinees. Also, the minimum values of SEs

[1]This was explained in Section 5.1.1.2 on page 69. This number was derived under the assumption that the standard deviation of the true abilities of the population is 1. In the simulations of Research Question 4, the standard deviation of the simulated true abilities were lower than 1. So, even though the dashed line gives some idea about the reliability of the test, a caution is advised when interpreting these results using this threshold.

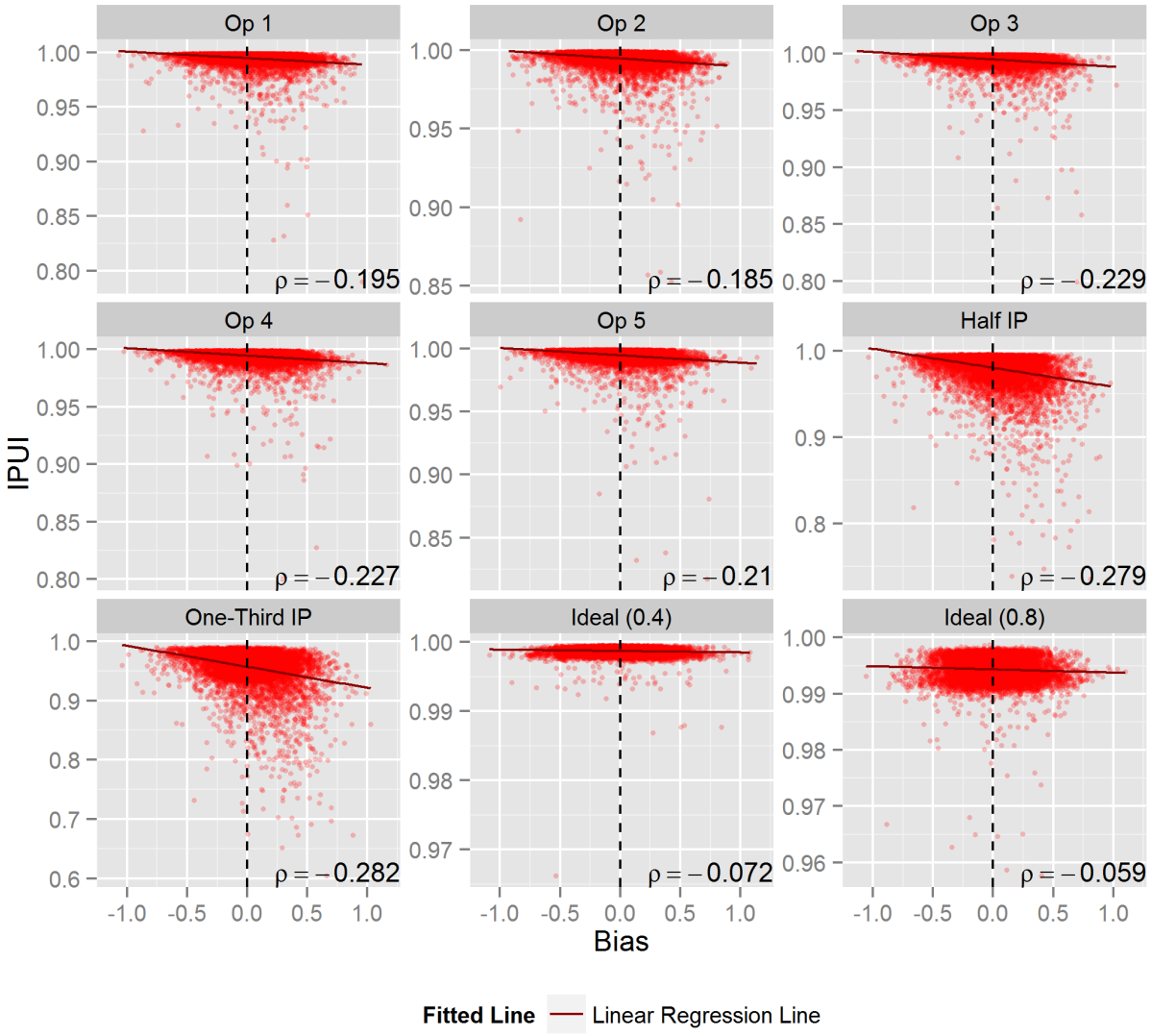

Figure 5.47: The Relationship between IPUI and Bias for each Item Pool Condition

for this item pool were larger compared to other item pool conditions. In this simulation, the examinees with small SEs were close to the cut score. For the examinees that were close to the cut score, SEs were larger for One-Third-IP condition. This can be interpreted as the weakness of that item pool.

Figure G.3 on page 192 shows the relationship between estimated ability and SE. There was a clear relationship between SE and estimated ability. Towards the extremes of the ability scale, the standard errors were higher. These were the examinees whose exams finished at

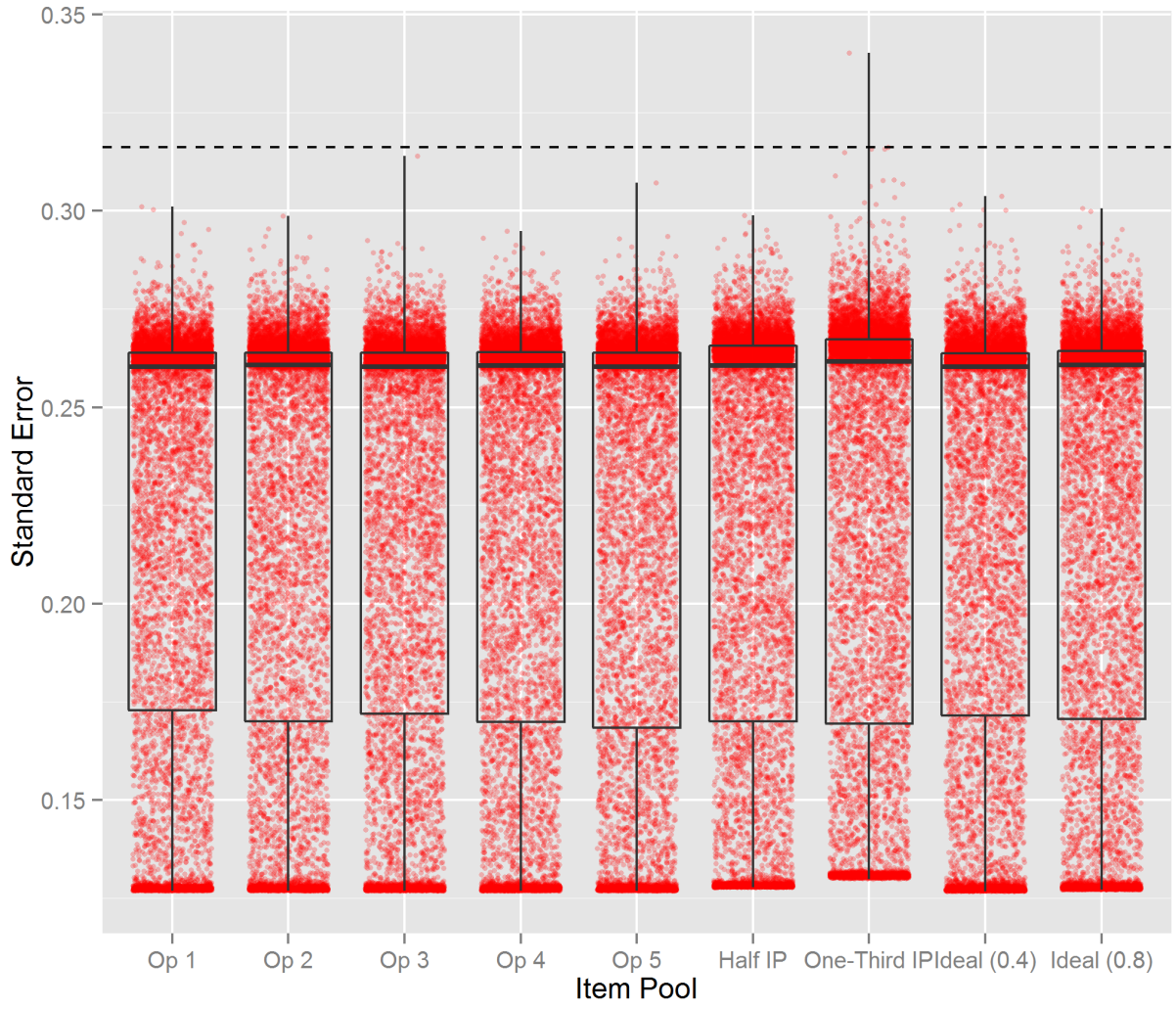

Figure 5.48: Standard Error Distribution for each Item Pool Condition

the 60th item. There was a steep decrease in SEs between $\hat{\theta} = -0.5$ and $\hat{\theta} = -0.25$, and steep increase between $\hat{\theta} = 0.25$ and $\hat{\theta} = 0.5$. Most of these SEs belong to the examinees with test lengths between 61 and 249. Lowest SE values between $\hat{\theta} = -0.25$ and $\hat{\theta} = 0.25$ belong to the examinees with test lengths 250. This pattern was the same for all of the item pools. But different than the other item pools, for One-Third-IP, the SEs increased even more at the extremes of the ability scale. Most probable cause of this increase was the scarcity of the items within the item pool. A similar trend can be seen to some extent for the Half-IP, especially for high ability examinees.

The relationship between SE and test length is shown in Figure G.4 on page 193. There was a clear negative relationship between test length and SE for all item pool conditions. The correlation coefficients in the text boxes show a very strong linear relationship between these two variables, even though the relationship appears to be curvilinear.

IPUI and Standard Error The relationship between SE and IPUI is shown in Figure 5.49. There is not an apparent relationship between SE and IPUI because this relationship is confounded with the test length. In fact, this figure resembles Figure 5.45, only the x-axis is flipped. There is a direct relationship between test length and SE as shown in Section 5.1.2.1 on page 77. For tests with 60 items, SE was high, IPUI was generally low for short tests as shown in Figure 5.45. As test length increased, the SE decreased. For operational and ideal item pools, the IPUI values were near 1 even for tests with 250 items. As a result, even though SEs decreased IPUI values remained close to 1. For One-Third-IP condition, IPUI values were lower when SE were lower. The reason for this was the lack of sufficient items in One-Third item pool to provide examinees with very long tests.

Mean Squared Error Figure 5.41 shows no difference between the average MSE values between item pools. The values in Table 5.7 also does not show any noticeable differences in either means or the standard deviations of the MSE values. Visual inspection of the individual MSE values in Figure 5.50 does not show any visible difference between item pools.

Figure G.5 on page 194 shows the relationship between IPUI and MSE. There was a weak negative association between these two variables. It was expected that the examinees who did not take the most appropriate items (i.e. had low IPUI values) also had high MSE values.

Exposure Rates NCLEX-RN is a high stakes test. Controlling the exposure rates is very important for the security of the test. Figure 5.51 shows the exposure rate distribution for each item pool condition. The dashed lines in the figure show the 0.20 and 0.05 levels for exposure rates, which corresponds to recommended high and low exposure thresholds

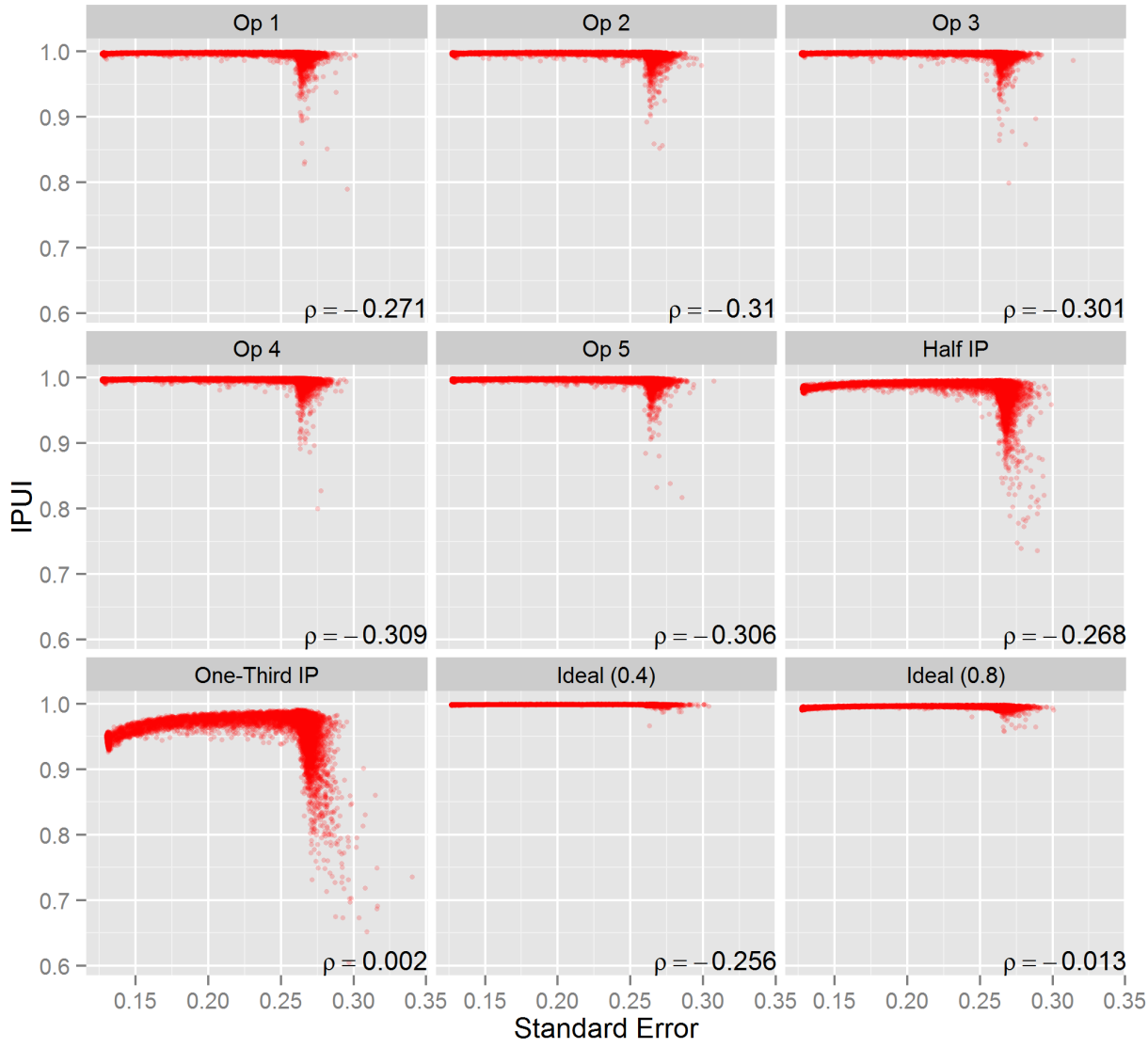

Figure 5.49: The Relationship between IPUI and Standard Error for each Item Pool Condition

for items, respectively. The exposure rate distributions of all operational item pools had a negative skew. Operational item pools were very successful in containing the exposure rates below 0.2. Nearly half of the items in operational item pools had exposures lower than 0.05. The median exposure rates (shown by the bold lines in the of the box plots) were all lower than 0.05. So, there were a lot of items that were administered to less than 5% of the examinee sample.

For the Half-IP and One-Third-IP, the exposure rate distributions had a balanced spread

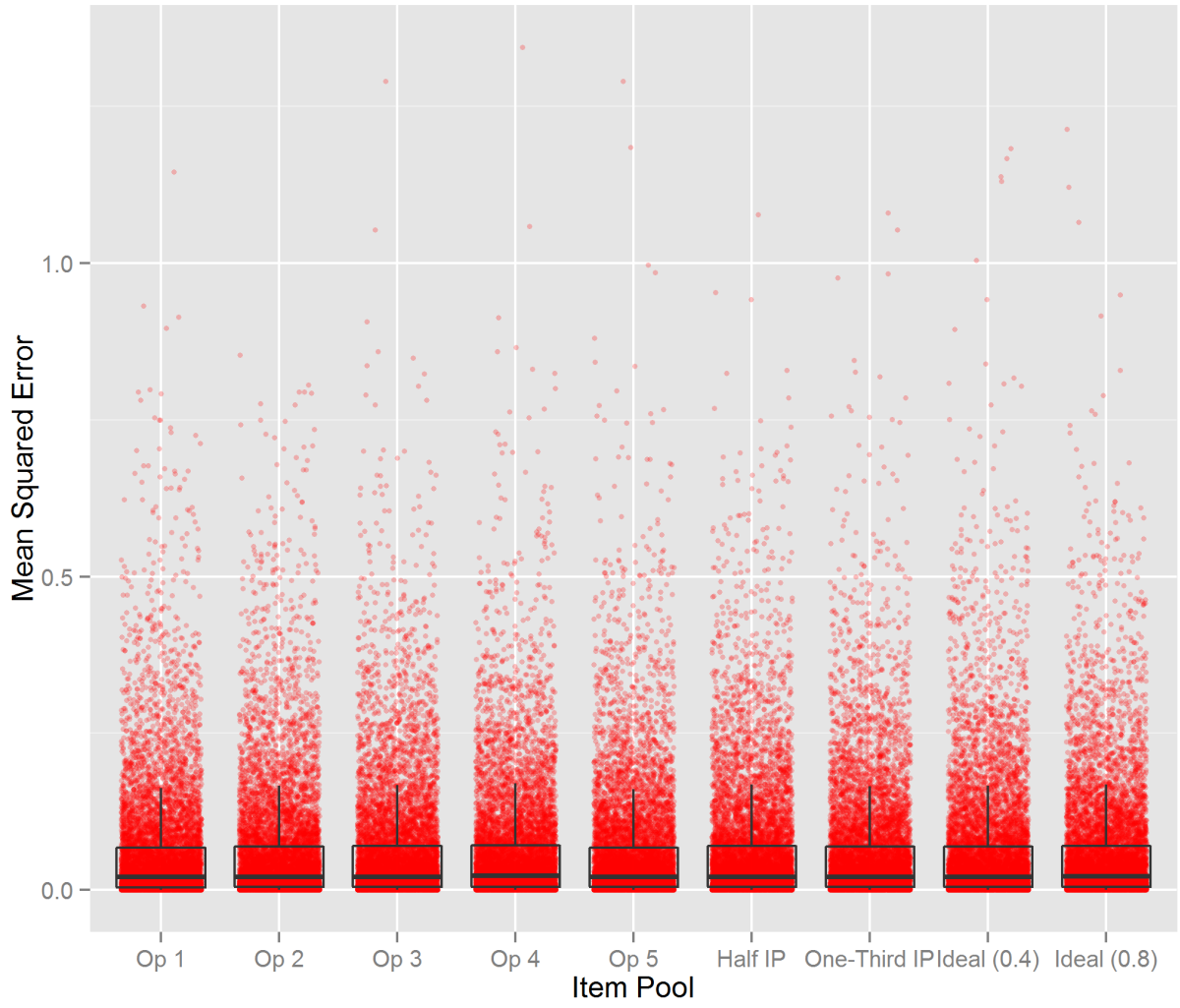

Figure 5.50: Mean Squared Error Distribution for each Item Pool Condition

and less skew compared to other item pool conditions. But many items had exposure rates larger than the 0.2 threshold. Especially for the One-Third-IP, the majority of the items had exposure rates larger than 0.2.

The performance of the ideal item pools were not perfect in the sense of exposure control. None of the items had exposure rates larger than 0.18 for Ideal-0.4 (ideal item pool with fixed bin size 0.4) item pool. But about 75% of the items had exposure rates lower than 0.05. For Ideal-0.8 item pool, many items were exposed to more than 20% of the examinees. But the majority of the items had exposures rates less than 0.05.

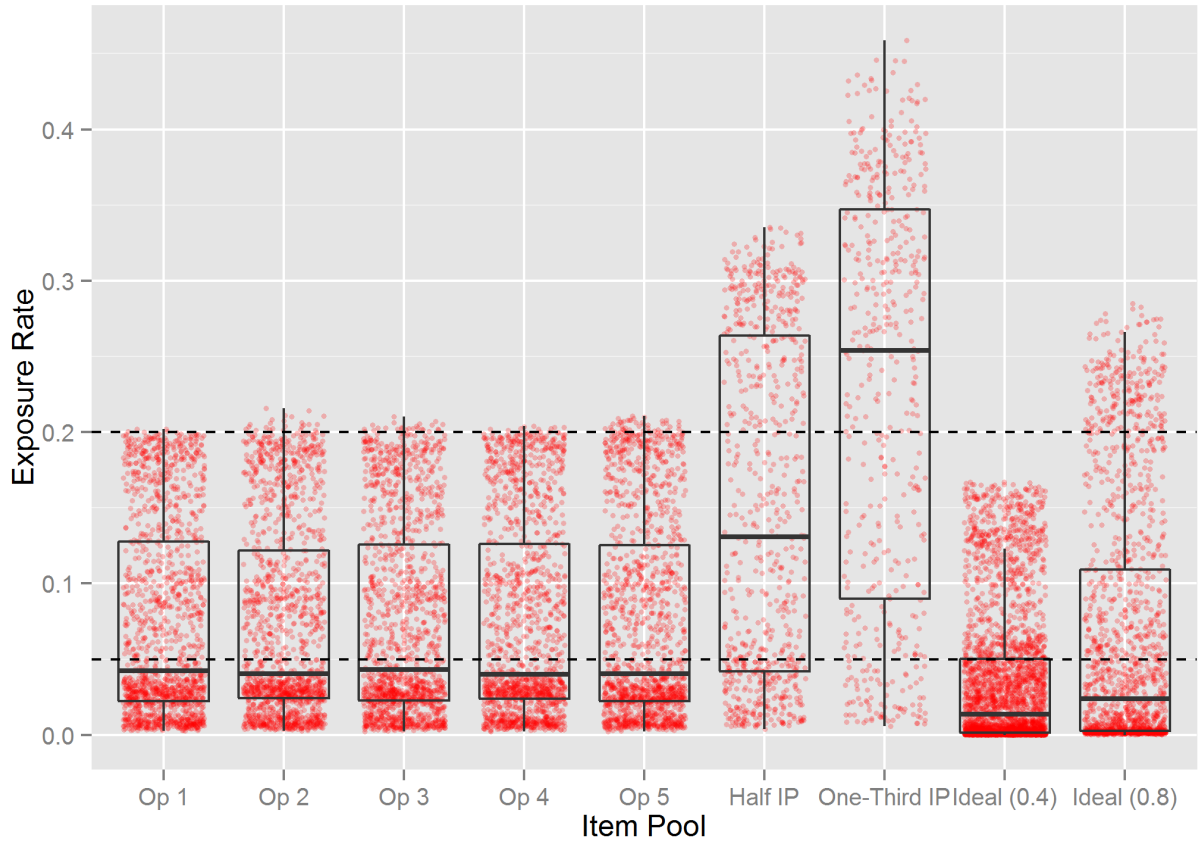

Figure 5.51: Exposure Rate Distribution for each Item Pool Condition

Table 5.8 shows the means and standard deviations of exposure rates in addition to the ratio of exposure rates that are larger than 0.20 and smaller than 0.05. Mean exposure rates reflects the number of items in the item pool. Because NCLEX-RN is a variable length test, there was not a functional relationship between mean exposure rates and item pool sizes as in the fixed test length tests. Exposure rate outcomes of operational item pools were very similar to each other. All had similar means and standard deviations for the exposure rates. Only a few items were exposed to more than 20% of the examinees. Within the operational item pools, Op-5 had the most items that had exposure rates larger than .2, which corresponds to 42 items (out of 1472 items). Almost half of the items were exposed to less than 5% of the examinees.

The results of the exposure rates were not perfect for any item pool. Operational item

pools successfully limit the exposure rates below 0.2 but they had a lot of underexposed items. The same is correct for Ideal-0.4 item pool. But this item pool had many more underexposed items (75% of the item pool). This item pool was not efficient in this sense. Half-IP and One-Third-IP conditions had fewer underexposed items, 31% and 16% of the item pools respectively. But they had many overexposed items as well, 39% and 59% of the item pools respectively.

Table 5.8: Item Exposure Analysis by Item Pool Condition

Item Pool	Mean Exposure	SD	Exposure > .20	Exposure < .05
Op 1	0.0737	0.0640	0.0034	0.5258
Op 2	0.0737	0.0641	0.0163	0.5326
Op 3	0.0738	0.0637	0.0170	0.5183
Op 4	0.0741	0.0648	0.0068	0.5312
Op 5	0.0745	0.0662	0.0285	0.5346
Half IP	0.1487	0.1108	0.3872	0.3139
One-Third IP	0.2254	0.1365	0.5869	0.1616
Ideal (0.4)	0.0354	0.0458	0.0000	0.7473
Ideal (0.8)	0.0662	0.0826	0.1362	0.6138

A further investigation of exposure rates revealed the problems for each item pool. Figure 5.52 shows the relationship between exposure rates and item difficulties grouped by content areas. For each operational item pool, most of the exposed items had difficulties close to the cut score. This is expected because the test lengths were longer around the cut score and the mean of the examinee population was close to the cut score as well. For operational item pools, most of the items that had low exposure rates were at the extremes of the ability scale.

One-Third-IP had most of the overexposed items around the cut score as well. But the peak point was around 0.3. There was a scarcity of mid-difficulty items. Most of the underexposed items for this item pool condition were at the lower end of the ability scale. Ideal-0.4 item pool had a lot of underexposed items at the extremes. None of the content areas had larger exposure rates compared to other content areas. One exception might be

the Ideal-0.8 IP condition. First content area had somewhat higher exposure rates in this item pool.

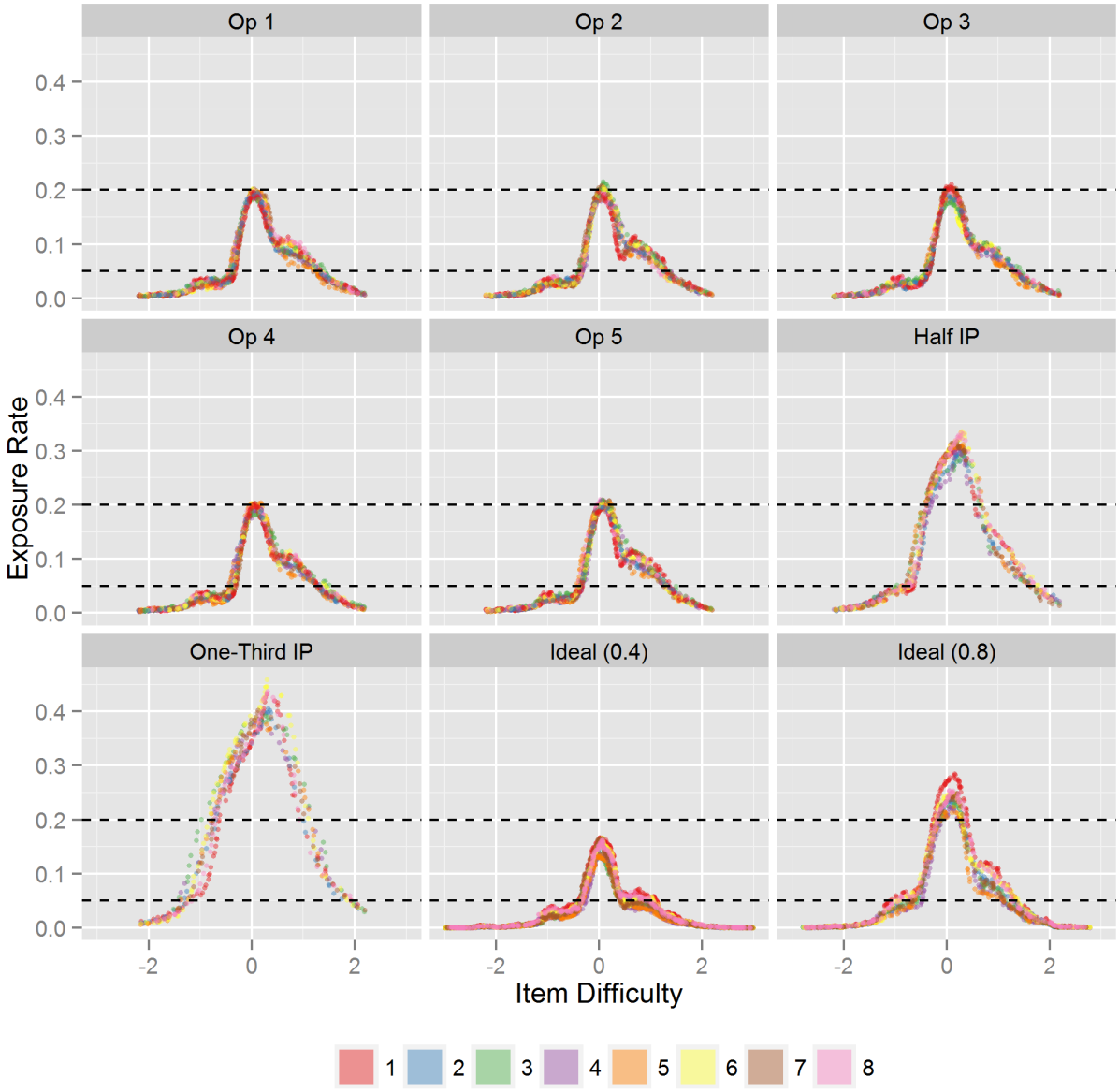

Figure 5.52: The Relationship between Exposure Rates and Item Difficulties Grouped by Content Area for each Item Pool Condition

Decision Accuracy NCLEX-RN is a certification test. So it is crucial to have a high decision accuracy. Table 5.7 shows the decision accuracy for tests based on each item pool.

For each item pool, decision accuracy was about 92%. Even the ideal item pool with over 3000 items did not have higher decision accuracy than the One-Third-IP. Table 5.9 shows the detailed information about the decision accuracy. Most of the examinees passed the test (around 66%). The decision was incorrect for about 8% of the examinees. The percentages of the false negatives (examinees who incorrectly failed) were higher than the false positives (examinees who passed incorrectly). Even though an incorrect decision is not good, usually for certification exams false negatives are better than false positives[2].

Table 5.9: Decision Accuracy Analysis by Item Pool Condition

Item Pool	Fail (C.D.)	Fail (I.D.)	Pass (C.D.)	Pass (I.D.)
Op 1	29.73%	4.15%	62.61%	3.51%
Op 2	29.64%	3.81%	62.95%	3.6%
Op 3	29.93%	4.08%	62.68%	3.31%
Op 4	29.77%	3.85%	62.91%	3.47%
Op 5	29.73%	4.06%	62.7%	3.51%
Half IP	29.69%	4.24%	62.52%	3.55%
One-Third IP	29.77%	4.3%	62.46%	3.47%
Ideal (0.4)	29.73%	4.27%	62.49%	3.51%
Ideal (0.8)	29.58%	4.01%	62.75%	3.66%

Percentages of simulees who failed or passed the test, and whether the decision was correct or incorrect.
C.D.: Correct Decision; I.D.: Incorrect Decision

5.2.4 Research Question 5

The aim of this research question is to demonstrate the diagnosis of the quality of an operational item pool using IPUI and guide test developers to build better item pools. Six item pools were investigated. Two of these item pools were operational item pools, two of them were ideal item pools with fixed bin sizes 0.4 and 0.8[3], two of them were one third and half of the first operational item pools (One-Third-IP and Half-IP respectively). For each

[2]In reality, this preference depends on the cost of a false positive and a false negative decision.

[3]Check Section 4.2.2.1 for the details of these item pools

item pool condition, the same NCLEX-RN CAT specifications as described in Section 4.2.2 on page 47 were used.

In the following paragraphs, item pools are diagnosed using different CAT outcome variables. The last outcome variable investigated is IPUI. It is hypothesized that IPUI can unearth the deficiencies in the item pools that other outcome variables could not find.

Bias The relationship between true θ and the mean of estimated θs for each item pool condition is shown in Figure H.1 on page 195. This graph shows two bumps close to the cut score. Away from the middle of the ability scale, there seems to be a perfect relationship between true θ and the mean of estimated θs. This graph does not show any difference between the item pool conditions.

Figure 5.53 shows the mean of biases at each true θ for item pools. This graph is a more detailed version of Figure H.1. Each item pool had a similar bump around the cut score. The reason for these bumps was the termination rule. The tests of the examinees ended when the cut score was outside the confidence intervals around their estimated abilities. Since their tests ended, examinees did not find the opportunity to converge to their true θs and consequently a bias occurred. For example, take the group of examinees with true $\theta = 0.25$. The mean biases for these examinees were 0.10, which means their mean estimated θs were about 0.35 ($= 0.10 + 0.25$). When these examinees correctly answered several items, their estimated θs increased (above their true θs) and since the cut score fell out of the confidence interval their tests ended. If the CAT algorithm administered more items to these examinees, their estimated θs would decrease. But since the test ended, it did not have the opportunity to converge on a good estimate.

Towards the extremes of the θ scale, there appears to be a difference between item pools. The ideal item pools (especially the Ideal-0.4 item pool) had mean biases close to 0. One-Third and Half item pools had negative bias close to θ values -3, and positive bias at θ values close to 3. The bias in One-Third IP is more visible for positive true θ values larger

than 2.

Figure 5.53: Mean Bias Conditional on True θ for each Item Pool Condition

Figure H.2 on page 196 shows the distribution of bias at each true θ value for the item pool conditions. Some of the true θ values close to the cut score were omitted to make the graph more readable. For all of the item pool conditions, the variation in the biases increased towards the extremes of the ability scale, except for the ideal item pools. The increase in the variation is more visible for One-Third item pool condition.

Standard Error Figure 5.54 shows the mean SE values at each true θ value for different item pool conditions. SE values close to the cut score were lower because of the long tests these examinees took. For true θ values outside of the θ scale between $[-1, 1]$, test lengths were almost always 60 (Figure H.5 on page 199). The difference between different item pools

become clear for these test lengths. One-Third item pool had the highest mean SE value followed by Half-IP. There was almost no difference between two operational item pools. The mean SE values for operational item pools were lower than Half-IP but higher than the ideal item pools. Ideal item pools had the lowest SEs along the θ scale. Ideal-0.4 item pool had smaller SEs even at the very extremes of the ability scale.

The results for the SEs outside the θ scale between $[-1, 1]$ reflects the number of items within each item pool around these values. Clearly lack of sufficient number of items inflated the SE values. Close to the middle of the ability scale, there was almost no difference between item pool conditions. At Figure 5.54, it is difficult to see the differences between item pools for true θ values close to the cut score. Figure H.3 on page 197 shows the mean SE for true θ values between -0.7 and 0.7. There was almost no systematic difference between item pools within this interval. One-Third-IP had slightly higher values between -0.2 and 0.2, but there was almost no difference outside this interval.

Figure H.4 on page 198 shows the standard error distribution at each true θ value for the item pool conditions. Some of the true θ values close to the cut score were also omitted in this graph to make it more readable. The variation of SEs was large at the middle and toward the extremes of the ability scale for the operational and reduced item pools. For the ideal item pools, the variation was large close to the middle of the ability scale, but variation did not increase towards the extremes. This can be seen as an indicator of the lack of appropriate items at the extremes for the operational and reduced item pools. The test lengths of the examinees at the extremes were 60. So, the only factor that affected the increase in SE should be the lack of appropriate items at the extremes.

Mean Squared Error Figure 5.55 shows the MSE values at each true θ value for all item pool conditions. Between true θ values -1 and 1, there appears almost no systematic difference between item pools. For true θ values around -0.4 and 0.4, MSE values made a dip. Close to true $\theta = 0$, MSE values reached a local maximum. Towards the extremes of the ability scale

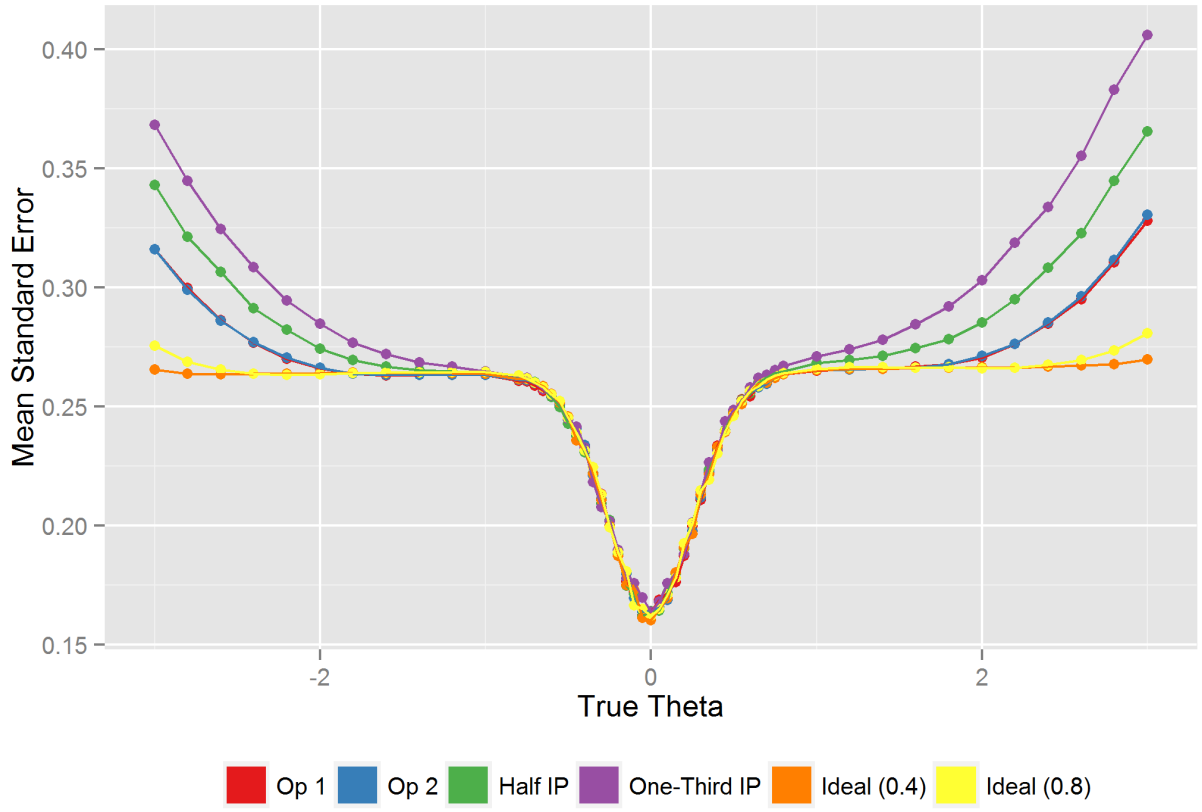

Figure 5.54: Mean Standard Error Conditional on True θ for each Item Pool Condition

the MSE values started to increase, except for the ideal item pools.

The systematic difference between item pools became clear for true θ values smaller than -2 and larger than 2. One-Third item pool had the largest MSE followed by the Half-IP. Two operational item pools had MSE values between the ideal item pools and the Half-IP. There was not a systematic difference between operational item pools. Ideal-0.4 item pool had the smallest MSE values. MSE values for this item pool did not increase even at the very extremes of the θ scale. Since MSE can be seen as a function of bias and SE this graphs makes sense. The difference towards the extremes were influenced by the differences in SEs among item pools (Figure 5.54).

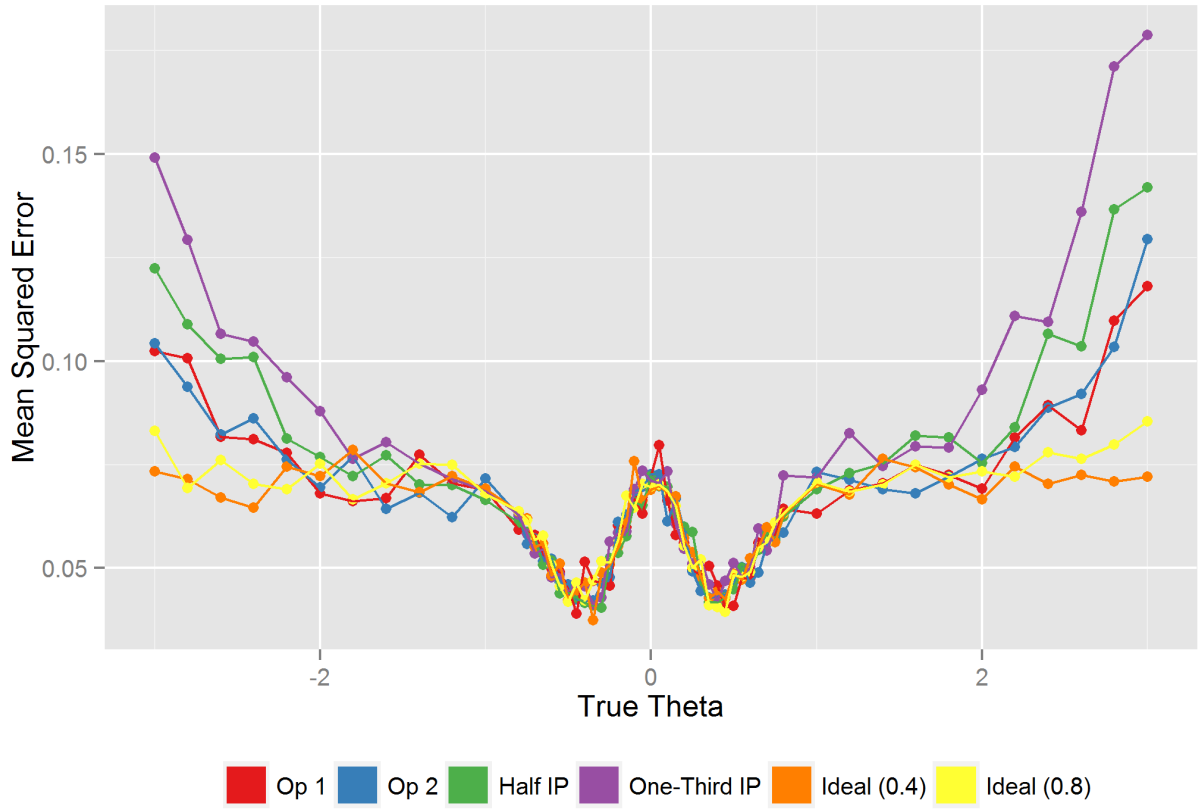

Figure 5.55: Mean Squared Error Conditional on True θ for each Item Pool Condition

Decision Accuracy Table 5.10 shows the decision accuracy at the true θ values between -0.5 and 0.5 for each item pool condition. The tabled results are restricted to this range because all the decision accuracy values outside this interval was virtually 100% for each item pool condition. Table 5.10 does not show any systematic difference between item pool conditions. Close to the cut score the decision accuracy was close to 50%. This make sense because if an examinee has true score that is equal to the cut score, then half of the time she will pass and a correct decision will obtained. As true θ values deviated from cut score, the decision accuracy improved.

IPUI Figure 5.56 shows the mean IPUI values at the true θ values for each item pool condition. This figure clearly shows differences between item pools. Each item pool had high

Table 5.10: Decision Accuracy Conditional on True θ for each Item Pool

True Theta	Op 1	Op 2	Half IP	One-Third IP	Ideal (0.4)	Ideal (0.8)
-0.50	100.0%	100.0%	99.9%	99.9%	99.9%	100.0%
-0.45	100.0%	99.9%	100.0%	99.8%	99.7%	100.0%
-0.40	99.8%	99.9%	100.0%	99.7%	99.4%	99.6%
-0.35	99.4%	99.4%	99.5%	99.5%	99.7%	99.3%
-0.30	99.1%	98.2%	98.7%	98.5%	98.7%	98.5%
-0.25	97.0%	97.9%	96.6%	95.3%	96.7%	96.9%
-0.20	91.7%	92.4%	91.9%	92.2%	93.1%	93.9%
-0.15	88.5%	84.8%	87.4%	87.0%	84.0%	86.4%
-0.10	78.1%	76.0%	80.7%	78.1%	75.7%	76.6%
-0.05	64.0%	64.6%	65.3%	64.8%	63.5%	63.8%
0.00	51.5%	47.4%	46.0%	51.4%	48.7%	48.8%
0.05	63.9%	63.9%	63.8%	65.7%	67.2%	65.2%
0.10	76.9%	80.3%	76.9%	78.1%	77.9%	76.9%
0.15	86.9%	87.5%	86.2%	88.1%	85.8%	85.1%
0.20	91.9%	93.3%	92.7%	93.8%	93.2%	93.8%
0.25	96.7%	97.0%	96.6%	96.5%	96.1%	96.4%
0.30	98.5%	98.6%	98.6%	97.9%	98.4%	98.5%
0.35	98.7%	99.2%	99.5%	99.3%	99.5%	99.3%
0.40	99.7%	99.9%	99.7%	99.8%	99.3%	99.7%
0.45	99.9%	99.9%	99.9%	99.8%	100.0%	100.0%
0.50	100.0%	100.0%	100.0%	100.0%	99.9%	99.9%

IPUI values close to the cut score. But for examinees that were away from the cut score, IPUI values started to decrease. Ideal-0.4 item pool performed the best. Throughout the θ scale, the values were close to 1 except very extreme θ values. Ideal-0.8 item pool had IPUI values above 0.99 between true θ values -2 and 2. IPUI values started to decrease towards the extremes.

Ideal-0.4 item pool shows the effect of lack of items for true θ values that were larger than 2.5 and smaller than -2.5. It is natural to ask why an ideal item pool does not have enough items for examinees at the extremes. The reason is simple. This item pool was ideal for a particular group of examinees, which had a distribution that resembles real examinees. Figure G.1 on page 190 shows the distribution of these examinees. None of the examinees in this distribution had true θs smaller than -2 and only a few were larger than 2. As a result,

the ideal item pools were not designed for the examinees outside this interval. It is normal for an examinee with true θ -3 or 3 to have an IPUI value smaller than 1 for these ideal item pools.

The performance of the two operational item pools were almost the same. Between true θs -1 and 1, the mean IPUI values were at or above .99. The mean IPUI values started to decrease towards the extremes. At $\theta = -2$ the value decreased to .95 and for $\theta = -3$ the mean IPUI value became .7. The performance of the operational item pools at the positive side of the θ scale was poorer compared to the negative side. At $\theta = 2$ the mean IPUI value reduced to .94 and at $\theta = 3$ it was .68.

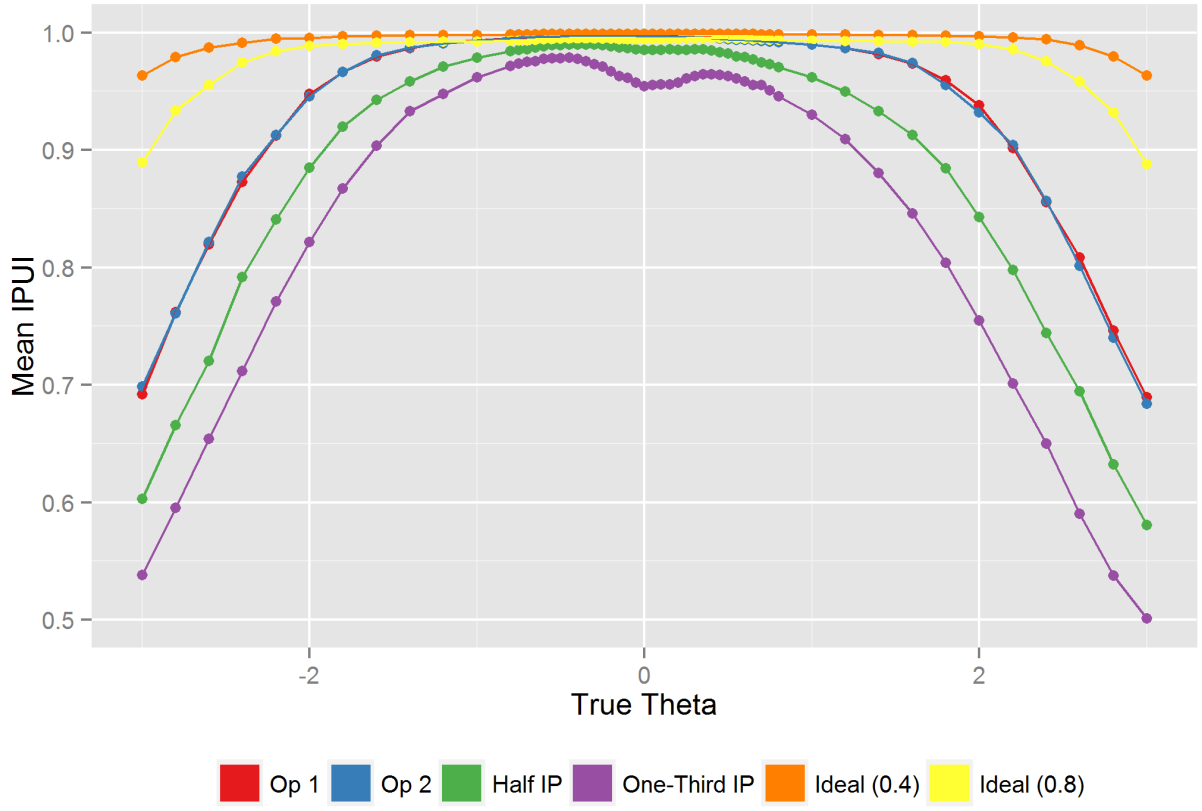

Figure 5.56: Mean IPUI Values Conditional on True θ for each Item Pool Condition

The performance of Half and One-Third item pools showed the same decreasing patterns towards the extremes of the ability scale. Even though the performances of item pools

between $\theta = -1$ and 1 are not distinguishable, the performance of One-Third item pool was clearly poor between this interval compared to the other item pools.

Figure 5.57 shows the mean IPUI values for item pools between $\theta = -0.7$ and 0.7. The difference between item pools are clear in this figure. Both operational and ideal item pools had IPUI values larger than 0.99 between this interval. Mean IPUI values for Ideal-0.4 item pool surpasses the remaining item pools throughout this interval. Operational item pools are indistinguishable in this figure as well. Ideal-0.8 item pool did not performed as well as the operational item pools close to the cut score. But as the previous figure showed, towards the extremes it's performance surpassed them. For none of the examinees Half item pool's IPUI values reach 0.99. This item pool did not perform comparatively well for the examinees at the positive side of the ability scale.

Even though Figure 5.57 shows a clear difference between the performances of item pools, the practical importance of this difference may not be large. This issue will be discussed in more detail in the discussion section. Here it can be said that IPUI shows even the slightest difference between the item pools that other CAT outcomes could not capture. Compared to the other outcome variables, IPUI clearly reflects the strengths and weaknesses of the item pools.

The results of the previous outcome variables were inconclusive. For example, at -0.10 the mean SE of the operational item pool 1 (Op-1) was lower than the remaining item pools, at -0.05 it was higher than the remaining ones. The same thing can be said for the bias and MSE. For none of these outcome variables can one derive a clear conclusion about whether an item pool is better than the others for a particular true θ value. IPUI can quantify the sufficiency of each item pool at each true θ value. Whether the differences between the IPUI values have a practical significance is another issue. The inconclusive results especially close to the cut score for bias, SE, MSE and the decision accuracy shows that the differences IPUI detected did not have practically significant effects on the other outcomes of the adaptive test.

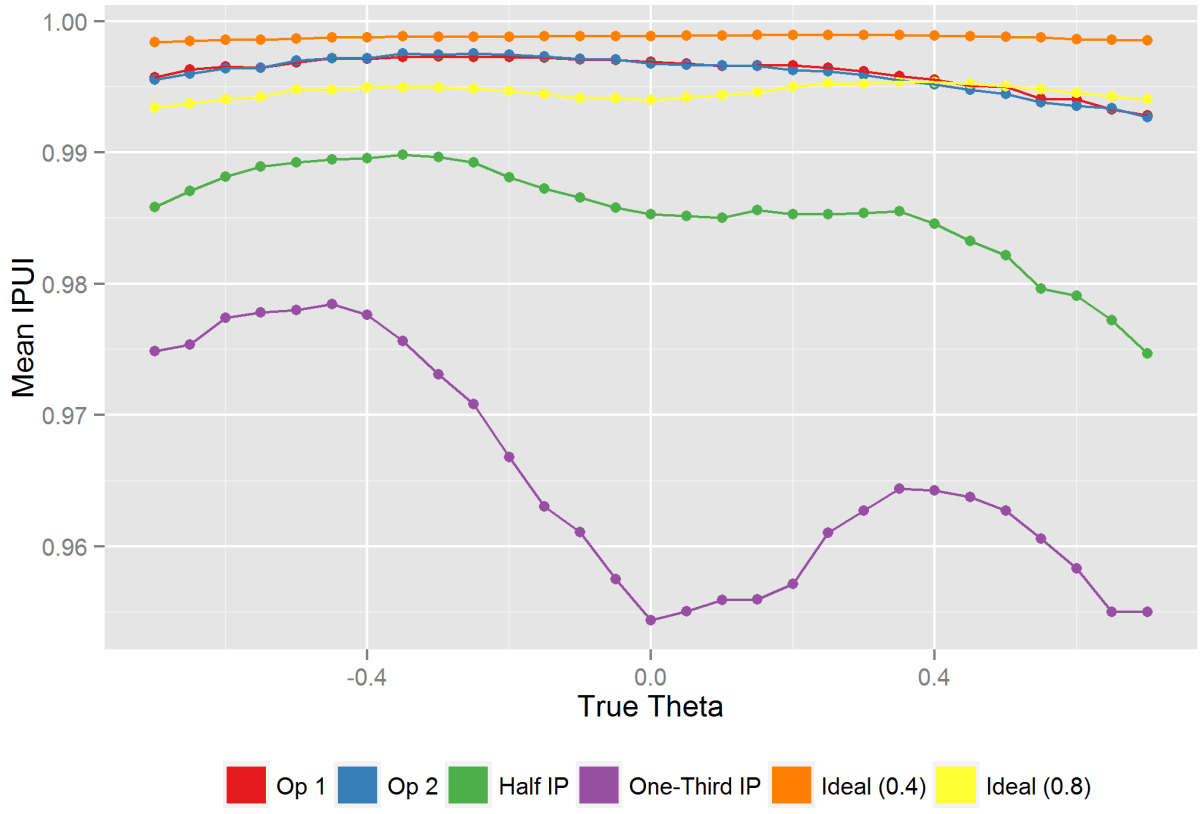

Figure 5.57: Mean IPUI Values Conditional on True θ around the Cut Score for each Item Pool

The previous two figures showed the mean values of IPUI. In addition to this, the distribution of IPUI also gives important information about the performances of the item pools. Figure 5.58 shows the IPUI distribution at each true θ value. Some true θ values around the cut score were omitted in this figure to make it more readable. The box plots shown for each true θ value include the median, quartiles and the spread of the IPUI distribution.

For each item pool condition, the spread of the IPUI values increased towards the extremes of the ability scale. Operational item pools performed well close to the cut score. Towards the extremes, the spread of the IPUI values increased for these item pools. The performance of reduced item pools were clearly worse than the operational item pools. The one-Third item pool did not performed well even close to the cut score. The spread of IPUI values were

large even for the examinees close to the cut score. The performance of Ideal-0.4 was the best. Even at the extremes, examinees can get the most appropriate items.

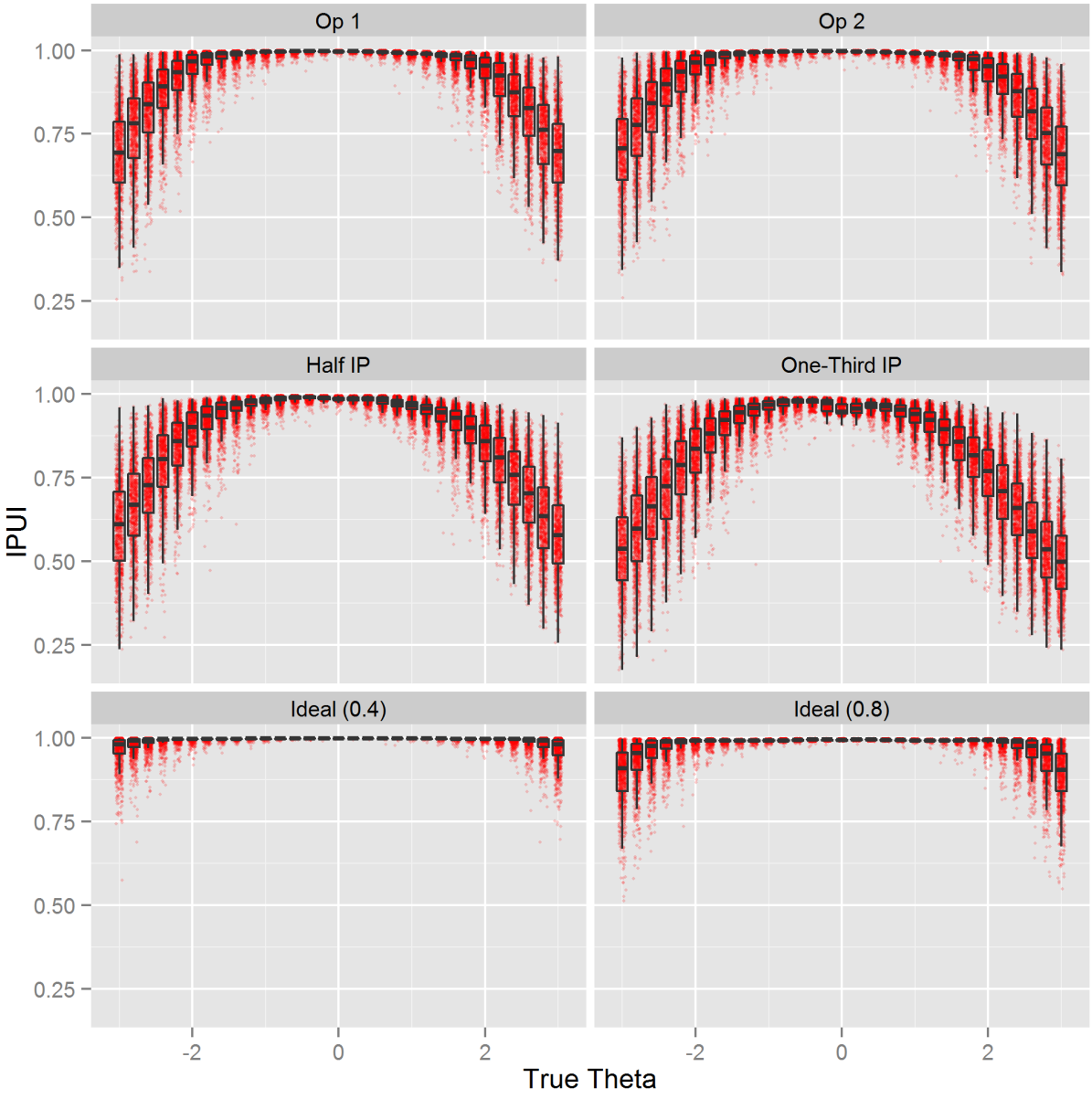

Figure 5.58: IPUI Distribution Conditional on True θ for each Item Pool

CHAPTER 6

DISCUSSION

6.1 Summary of the Results

The aim of this study was to develop a method to evaluate the quality of item pools for adaptive tests. Current methods to evaluate the adequacy of the item pools were discussed in Section 2.3.6. These methods fall short as the CAT test design gets complicated. The histogram of the item parameters for an item pool might show that item pool is sufficient. But changing a specification of the CAT test might make this item pool insufficient for the purpose of the test. To solve this problem, a new index evaluating the quality of item pools was developed. This index is called Item Pool Utilization Index (IPUI, see Section 3.2 on page 27 for the derivation of this index). IPUI quantifies the amount of deviance of an item pool from a perfectly optimum item pool. This theoretical optimum item pool satisfies each specification of the test and provides an optimum item in terms of information at each stage of the test for every examinee.

The utility of this newly developed index was investigated by five research questions (see Section 4.1 on page 36). The first three research questions used simulated data and the last two questions used operational data.

Research Question 1 investigated whether IPUI is sensitive to the changes in the quality of the item pools. Item pool quality was operationalized by (1) the discrepancy between the item difficulty distribution of the item pool and the ability distribution of the examinees, and (2) the item pool size. For the first part of the Research Question 1, 13 discrepancy conditions were tested. The item pool was the same for all conditions. Item difficulty parameters of the item pool were generated from the standard normal distribution. The examinee ability distributions were generated from a normal distribution. The means of the distributions

ranged from -3 to 3 with 0.5 intervals. The results (Figures 5.1 and 5.3 on page 54 and on page 58) showed that IPUI was sensitive to the discrepancy between item pools and ability distribution. Increasing discrepancy affected the other outcomes of the adaptive test as well, such as SE. But IPUI had more variability in its evaluation of the shortcomings of item pool compared to SE, and the relationship between SE and IPUI was not linear (Figure 5.4 on page 59).

Second part of the Research Question 1 investigated whether IPUI is sensitive to the changes of the size of an item pool. Eleven different item pools ranging from 20 items to 1000 items were investigated. As the number of items in the item pool increased, the mean value of IPUI increased as well (Table 5.2 on page 67). After 300 items, the increase in IPUI values were minimal.

Additionally, Research Question 2 looked at the sampling distribution of IPUI. For each item pool size condition, 25 replications were performed. Within each condition, each replication had the same item pool size. The item difficulty parameters were generated from the same distribution. Table 6.1 shows the mean and standard deviation of mean IPUI values aggregated by replication. The variability of the mean IPUI values were larger for small item pool sizes. The variations of mean IPUI values had two sources. First, the variation due to sampling of different item pools with the same sizes for each condition. Second, the variation due to IPUI.

The specifications of CATs can get rather complex. An item pool that is working perfectly fine for one set of specifications might not perform in a similar way for another set of specifications. Research Question 2 investigated whether IPUI can detect the adequacy of the same item pool for different CAT specifications. Two CAT specifications were investigated: test length and exposure control.

First part of the Research Question 2 investigated 18 different test lengths ranging from 5 to 400[1]. Results showed that as test length increased the value of IPUI decreased (Figure 5.20

[1]The size of the item pool was also 400

Table 6.1: Means and Standard Deviations of Mean IPUIs of the Replications

Item Pool Size	Mean of Mean IPUI	Standard Deviation of Mean IPUI
20	0.5875	0.02528
40	0.8348	0.00935
60	0.9028	0.00803
80	0.9345	0.00928
100	0.9494	0.00843
200	0.9793	0.00448
300	0.9880	0.00210
400	0.9906	0.00217
500	0.9924	0.00145
750	0.9953	0.00090
1000	0.9965	0.00096

on page 83). IPUI distributions indicated that this item pool can support a test length of 50 for majority (75%) of the examinees.

Second part of the Research Question 2 investigated 12 exposure control conditions. These conditions range from no exposure control to random selection of items. IPUI detected even small differences between conditions where other CAT outcomes showed no difference (Figure 5.22 on page 87). Results of the Research Question 2 showed that IPUI is very sensitive to even small modifications of the test specifications.

Research Question 3 was designed to show the utility of IPUI as a diagnostic tool for item pool evaluation. Three test plans with different specifications were investigated. The first two plans had the same item pool but in the first plan content balancing was imposed on item selection. For the second condition, there was not any constraints on item selection algorithm. The third plan had an item pool consisting of rather difficult items. IPUI results clearly showed the weak points of the item pool for each condition. Further diagnostic information provided by different graphs of IPUI showed detailed information of the weaknesses of the item pool for particular test specifications.

Research Question 4 and 5 used operational data provided by NCSBN to show the utility of IPUI. Research Question 4 compared nine different item pools with the same specifications

as the NCLEX-RN exam. Five of the item pools were the operational item pools previously used in NCLEX-RN exams. Two item pools were the ideal item pools generated for the specifications of the NCLEX-RN exam. Two item pools were generated by randomly removing half and one-third of the first operational item pool. The same examinee distribution as the real examinee population was used for the comparisons. The results of the Research Question 4 showed that all of the the item pool designs performed well for the examinee group. Among other outcomes of CAT only the IPUI detected the weaknesses of the half and one-third item pools (Figure 5.41 on page 113). IPUI detected that these two item pools depleted of appropriate items towards the end of the test for examinees close to the cut score (see Figure 5.45 on page 119). Operational and ideal item pools were strong even for the long tests.

Research Question 5 was similar to the Research Question 3. The aim was to show the utility of IPUI as a diagnostic tool for an operational CAT. The results showed that operational and ideal item pools were very robust close to the cut score. Towards the extremes of the ability scale, the operational item pools started to weaken (see Figure 5.56 on page 137). But this had no effect on the decision accuracy (see Table 5.10 on page 136). IPUI detected even slight differences between item pools close to the cut score where each item pool was comparatively strong (see Figure 5.57 on page 139).

6.2 Practical Uses of IPUI

The results of the study showed that IPUI was very sensitive to the changes in the quality of item pools, changes in test specifications that affected the utilization of the item pool, and was a useful diagnostic tool to improve the item pool quality. In this section the practical uses of IPUI are discussed.

6.2.1 Quantification of the Item Pool Quality

Test developers are building item pools for adaptive tests very frequently. Users of these tests need to know the quality of the tests that are administered. Since the quality of the adaptive tests are strongly tied with the quality of the item pools (Flaugher, 2000), test developers have to make sure that the quality of their item pools are adequate for the purposes of the test.

Four general methods are used to get information about the item pool: (1) item pool size, (2) descriptive statistics for item pool parameters (i.e. mean, standard deviation, histogram, etc.) (3) item pool information function (4) outcome variables of CAT simulations. Each of these existing methods has their own shortcomings as explained below.

Item pool size, i.e. the number of items in the item pool, is an important indicator of the adequacy of the item pool. Even though general rules exist for the size of an item pool, such as item pool size should be twelve times the length of the adaptive test (Stocking, 1994), there are no one size fits all kind of general rule for an adequate size for an item pool. An item pool which is perfectly adequate for an examinee group might not be adequate for another examinee group (see the first part of the Research Question 1 in Section 5.1.1.1 on page 53). A high quality 100 items might perform better than a low quality 200 items (Xing & Hambleton, 2004). As a result, in addition to the size of the item pool, the information about the quality of the items are necessary to evaluate the item pool. The quality of items can be measured by the item parameters.

The descriptive statistics for the item pool parameters are useful to see the overall picture of the item pool. Common descriptive statistics are the means and standard deviations of the item parameters or the histograms which are used to visualize the item parameters. Especially the distribution of the item difficulty parameters can be helpful to see whether there is a discrepancy between the item pool and the ability distribution of the examinee group. For instance, a visual comparison of Figures A.1 and A.2 on page 167 and on page 168

will give an idea about the ability distribution to which the item pool is most appropriate. But operational CATs are rarely as straightforward as the ones in Research Question 1. There are many constraints on the item selection algorithm which makes it difficult to evaluate the sufficiency of the item pool by simply inspecting the descriptive statistics of the item parameters. For example, plan 1 and 2 in the Research Question 3 used the same item pools (Figure 5.30 on page 98). But due to the content balancing imposed on the item selection algorithm in plan 1, the performances of these two item pools differed a lot (see Figures 5.32 and 5.34 on page 100 and on page 102).

Item pool information functions are widely used to evaluate the adequacy of the item pool for a particular test purpose. Xing and Hambleton (2004) used item pool information functions to compare different item pools (see Figure 2.1 on page 24). In addition, item pool information functions are widely used to build item pools for adaptive tests (van der Linden, Veldkamp, & Reese, 2000, 2006; Belov & Armstrong, 2009). But test information functions suffer from the same disadvantage explained in the previous paragraph. Two item pools might have the same information functions (as in plan 1 and 2 of Research Question 3) but they can perform differently.

The comparison of the item pools using the outcomes of the CAT, such as biases, SEs of the ability estimates and the exposure rates of the items is a common approach as well (He & Reckase, 2013; Thompson & Weiss, 2011). In fact, in all of the research questions investigated, such outcomes of CATs were used along with the IPUI to compare the item pools. The outcomes of the CAT gives valuable information about the performance of the item pools. But none of them gives a direct way to evaluate the quality of an item pool. Figure 1.1 on page 2 is a very good example for this. This figures shows that the item pool provided very appropriate items to the Examinee 3, but not to the Examinee 4. The IPUI values of Examinee 3 and 4 was 0.98 and 0.273 respectively. But the SE of the Examinee 3 was higher than the SE of Examinee 4. If one judges the performances of the item pool for these two examinees according to SE, an inaccurate picture can be drawn. Such examples

can be found for the other outcomes of the CAT.

As it was shown in this study, IPUI is a direct way to measure the adequacy of an item pool at both the examinee level and the examinee group level. A test developer can easily quantify the adequacy of item pool by using IPUI without resorting to the indirect ways to evaluate the item pools.

6.2.2 IPUI in Optimal Test Assembly

The special issue of Applied Psychological Measurement in September 1998 was about the optimal test assembly. van der Linden (1998b) introduced the concept and discussed different methods to assemble a test optimally. He divided the test specifications into two broad areas, constraints and objectives. Constraints are the test or item attributes that has an upper and/or lower limit to be met. For example, a minimum or maximum number of items to be administered, the number of words in the test, the number of items with figures and etc. are among the constraints. The objectives require a test attribute or a function of item attributes to reach a minimum or maximum. For example, maximization of test information within a certain range, maximization of the test validity, maximization of the decision accuracy, minimization of the standard error of ability estimates and etc.

These constraints and objectives generally lead to an optimization problem: optimization of an objective in the presence of the constraints. IPUI can be used as a constraint or as an objective in these optimization problems. As a constraint, a test developer might require that none of the examinees in an adaptive test should have an IPUI value less than a certain value. On the other hand, IPUI can serve as an objective of a test assembly problem to be maximized. An item pool that has a maximum mean IPUI value that satisfies all of the constraints of the test can be chosen as an item pool.

6.2.3 IPUI as a Quality Control Tool

Tests, especially the high stakes ones, should conform to industry standards as described in Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Test developers constantly monitor the quality of the tests they administer. In addition to adhering to the high quality test development techniques (Schmeiser & Welch, 2006), developers of the CAT test should make sure that their tests are performing as intended.

For CAT tests, test developers have less control over the particular test an examinee gets. Hence, the need for quality control is high. One of the methods test developers use is showing the paper-and-pencil copies of various CAT tests to expert test specialists (Eignor et al., 1993). These experts examine the tests and check whether the tests adhere to the test specifications.

IPUI can be used as an additional tool for checking the quality of individual tests that are administered to the examinees. If an examinee's IPUI value drops below a certain level, the examinee might be flagged. For instance, Figure 5.43 on page 116 showed that for One-Third item pool, examinees with ability estimates larger than 2 had relatively lower IPUI values. If this is not acceptable considering the test purpose, test developer can take appropriate precautions. A low IPUI value signifies that item pool failed to provide appropriate items to this examinee. Test developer can either improve the item pool or change the test specifications to improve the utilization of the item pool.

6.2.4 IPUI as a Diagnostic Tool

As indicated in the previous sections, IPUI can be used to measure the quality of the item pool and as a tool of quality control. After finding out that item pool is not performing well, the next step of a test developer is to diagnose the deficiencies of the item pool and improve the item pool in such a way that item pool provides each examinee appropriate items.

Diagnostic utility of the IPUI was investigated in Research Questions 3 and 5. The results of these research questions showed that IPUI can show the deficiencies of the item pools and guide test developers to add suitable items to fix these deficiencies. For example, in Research Question 3, it is difficult to judge from the SE graph in Figure 5.32 on page 100 the properties of the items that should be added to the item pool in plan 1. The IPUI vs. item number graph in Figure 5.35 on page 104 showed that the cause of the low performance of this item pool was the content balancing restrictions. Further, Figure 5.36 on page 105 showed the approximate item difficulty values needed from each content area to improve the item pool. Similar graphs that uses other test constraints can be helpful to diagnose and improve the item pools.

6.2.5 IPUI at Individual and Group Level

The previous sections explained four different ways the IPUI can be used in practice. It is believed that IPUI can be used as an outcome variable for CAT just like bias, SE, MSE, exposure rate or overlap rate. Test developers can use IPUI at two levels: at group level or at individual level.

At group level, IPUI is an indicator of the adequacy of an item pool for a given set of test specifications and examinee group. The group level statistics can be the mean or median of the IPUI. These statistics will influenced by the examinee group. If the item pool is appropriate for most of the examinees the mean of IPUI will be large. If the item pool is appropriate for only a small portion of the examinees tested, the mean of IPUI will be small.

Practitioners can use these summary values to evaluate the quality of their item pools over time. If the examinee group does not change dramatically from year to year, test developer can set a standard minimum for the mean IPUI value. Over the years each item pool developed can be compared to this standard. This will ensure the fairness of the test across years. Research Question 2 showed that a change in test specifications can affect the adequacy of the item pool. Using a standard like this will allow test developers to implement

new test specifications while ensuring that the adequacy of the item pool is still on par with the item pools used in the previous testing windows. For instance, if the testing agency decides to increase the security of the item pool by implementing a new exposure control method, test developers can use IPUI to ensure that the item pool is still adequate after such a change.

A second use of IPUI is at the individual level. A test developer can set a minimum IPUI value for each examinee so that the item pool is adequate for each test taker. In practice, the item pools cannot provide appropriate items to the examinees at the extremes of the ability scale. For instance, operational item pools in Research Question 4 did not provide appropriate items to some examinees with estimated θ values smaller than -2 or larger than 2 (Figure 5.44). For this operational test, the inadequacy of the item pool for these examinees did not affect the decision about the examinees. On the other hand, if the purpose of the test was to measure each student sufficiently well, the test developer can make the item pool broad enough to provide appropriate items to the examinees at the extremes. This will ensure the test fairness at individual level. An item pool that provides appropriate items to some group of examinees but not for another group will undermine the fairness of the test. Using a minimum value for IPUI as a benchmark, test developers can ensure the quality of service for each individual examinee.

From the perspective of an examinee, they want a fair instrument that measures their ability as precisely as possible. If the examinee gets a test with high IPUI value, this means at least the item pool portion of the CAT worked fine.

As discussed in Section 6.4.2 on page 156, this study does not provide a recommended value for IPUI. But this does not preclude test developers to set their own standards and compare the performances of the item pools using IPUI.

6.3 Implications

6.3.1 The Robustness of CAT Procedures to Weak Item Pools

The results of the study showed that IPUI was very sensitive to changes in the quality of the item pool. When a CAT algorithm administers sub-optimal items, IPUI detects them. The other outcomes of the CAT were not as sensitive to sub-optimality of the item pool unless the item pool underperformed significantly. For example, mean bias was rarely affected by the inadequacy of the item pool unless there was a large discrepancy between item pool and ability distribution (see Figure 5.31 on page 99).

In Research Question 4, the fidelity coefficient, the mean bias, the decision accuracy and the mean SE values were almost the same across the different item pool conditions (see Table 5.7 on page 114). Even though IPUI indicated a performance difference among item pools, this did not reflect on the other outcomes. The results of the other research questions indicated this too. This reflects the robustness of the CAT procedures to inadequate item pools.

The robustness of maximum likelihood ability estimation in a CAT was shown by Chang and Ying (2009). They found that even for an item bank with a limited capacity, the maximum likelihood estimates of θ were consistent and asymptotically normal.

For the Rasch model, the robustness of the CAT procedures to the selection of sub-optimal items was observed by other researchers too. Bergstrom et al. (1992) performed a study where they modified the item selection algorithm to select items with 0.5, 0.6 and 0.7 probabilities of correct responses. The effect of these modifications on the precision of the ability estimates was minimal. Way (1998) concluded that "...the adaptive nature of CAT plays a surprisingly minor role when the Rasch model is used. This contradicts the commonly held assumption that CAT will significantly improve measurement precision through targeting item selection to each individual" (p. 21). The results of this study corroborates the findings of these researchers.

6.3.2 Summary Statistics for IPUI

When evaluating the IPUI for a group of examinees, a test developer has different options to summarize the distribution of IPUI. The most informative way is to visualize the distribution of IPUI values using a histogram, box plot or a scatter plot of the IPUI versus the θ estimates[2]. This will allow the test developer to observe the performance of the item pool at an individual level.

At the group level, the definition of IPUI in Equation (3.6) on page 29 uses the mean to summarize the performance of the item pool. Especially for skewed IPUI distributions, averaging the IPUI values might not give a good picture of the adequacy of the test. The distribution of IPUI is generally negatively skewed if the item pool is well suited for the majority of the examinees. This can be seen for some conditions in the results section (i.e. Figures 5.3 and 5.20 on page 58 and on page 83). In such cases, the mean and median values of IPUI differ and may lead to different interpretations about the quality of the item pool[3]. This will influence the potential comparisons of the item pools. When the mean and median values of IPUI are discrepant, practitioners are advised to look at the overall distribution of the IPUI values and evaluate the item pools accordingly.

In addition to the mean or median values of IPUI, the variation of the IPUI can give important pieces of information. The best case scenario for an item pool is a high mean and a small standard deviation of IPUI values. This happens when the item pool provides appropriate items to almost all of the examinees. However, if the variation of the IPUI values is large, this indicates large discrepancies between the performance of the item pool for different examinees. Depending on the purpose of the test, this might not be fair.

[2]See Figure 5.44 on page 118 as an example.

[3]See the end of Section 5.1.2.1 on page 82 for an example and a discussion about the differences between using median and mean as a summary statistics for IPUI.

6.3.3 Commentary on the Results of the Operational Item Pools

The results for the comparison of the operational item pools were very good. Except for the examinees who were far away from the cut score, performances of the operational item pools were very good. Since the purpose of the test was dividing examinees into two groups, the effectiveness of the item pool at the extremes might not be crucial. Operational item pools performed very well around the middle of the ability distribution where the cut score was located. Examinees who were closer to the cut score took longer tests. So, it was important for item pool to provide sufficient number of appropriate items for the examinees close to the cut score. The graphs comparing IPUI and test length[4] showed that the IPUI values of the examinees who took long tests were very high. This suggests that the operational item pools provided appropriate items to the examinees whose tests lasted 250 items. These examinees were the ones for whom the measurement precision was very essential.

In fact, NCLEX-RN exam was not a very good example to show the merits of the IPUI. The exam has a long history and the item pools for the operational tests are meticulously prepared for the test. Furthermore, the test is very long, consequently it is very hard to get a decision error unless the examinee's true score is close to the cut score. The merits of IPUI would be more evident for tests with much smaller item pools and the decision for a wide range of abilities are needed. Achievement tests which desire to measure a wide range of abilities would be a good example for showing the uses of IPUI.

6.3.4 IPUI and Measurement Quality

IPUI is an indicator of the adequacy of the item pool. It is not an indicator of the quality of the measurement. Certainly, an adequate item pool would improve the quality of the measurement, but it is not a sufficient condition. An item pool might provide appropriate items to an examinee, but still, the measurement quality of the test might be low.

[4]See Figure 5.45 on page 119.

For example, Figure 3.2 on page 32 shows that the item pool provided appropriate items to Examinee 3 throughout the test. The IPUI value for this examinee was 0.98, indicating the item pool was adequate for this examinee. But the SE of the ability estimate for this examinee was high, 0.685. The test ended after 8 items, and for this examinee, 8 items were not enough for a precise estimate of the ability. Even though the item pool portion of this test performed well, the test specifications needs to be changed for a precise measurement of the ability.

On the other hand, a precise ability estimate does not mean that the item pool performed well. For example, Examinee 4 in Figure 3.2 had lower SE compared to Examinee 3, but the item pool failed to provide appropriate items to this examinee. The results of the first part of the Research Question 2 also corroborates this. Figure 5.14 on page 74 shows that when tests were longer the ability estimates were more precise. Yet, the item pool failed to provide enough appropriate items for longer tests.

There are other situations where an item pool is adequate but due to the other aspects of the CAT algorithm, the measurement quality suffers. For instance, if the item selection algorithm (such as EAP or MAP) uses a strong prior distribution, the ability estimate will be biased (Kim & Nicewander, 1993). The item pool might provide appropriate items to the examinees, but this will not reduce the bias caused by the item selection algorithm. A high IPUI value might not correspond to smaller biases.

The results of this study showed that, in general, an adequate item pool is associated with better measurement[5]. Nevertheless, as discussed, an adequate item pool does not always enough for high measurement quality.

[5]See Figures 1.1, 5.4 and 5.11 on page 2, on page 59 and on page 70.

6.4 Limitations of the Study

6.4.1 Generalizability of the Results

The generalizations made in the study are limited to the methods used. For example, in the second part of the Research Question 2, the effect of exposure control on IPUI was investigated. In that research question, only the randomesque exposure control method was used as a proxy of exposure control methods. As discussed in Section 2.3.3.2 on page 14, there are many other exposure control methods used in operational tests. The results presented in this study are limited to randomesque exposure control method. But still, it is expected that any exposure control method will reduce the quality of the item pool. Similar limitations of generalizability are also valid for the other aspects of the simulations.

The item selection procedure is another example of the limited generalizability of the current study. MFI item selection algorithm was used in all of the simulations in this study. Results might be different for other item selection algorithms. The generalizability of the results from MFI to other item selection algorithms might not be straightforward. MFI uses the Fisher information of the items. The IPUI also depends on the Fisher information. In this respect, IPUI is very relevant to CATs using MFI. On the other hand, other item selection algorithms might have a different criteria for selecting the items. For example, Kullback-Leibler item selection algorithm (Chang & Ying, 1996) searches for an item that maximizes the global information instead of Fisher information. As a result, even if there are no other constraints on the item selection algorithm (such as exposure control or content balancing) and the item pool is sufficiently large, the IPUI value might be low for this item selection algorithm. In this regard, this might be seen as a limitation of IPUI. But IPUI can be generalized and reformulated such that instead of maximization of the Fisher information, maximization of the global information might be expected[6].

Operational CATs have many other constraints such as item enemies, limitations on the

[6]More on this in Section 6.5.1 on page 161

number of certain types of items, limitations on word counts, key distribution constraints (i.e. same number of A's, B's ... should appear as correct response.) and etc[7]. These were not explored in this study. Each of these constraints are expected to reduce the value of the IPUI.

Throughout this study, the effect of changing only one CAT specification on IPUI was investigated. In Research Question 2, as the test length increased -while holding every other aspect of the CAT- the mean value of IPUI decreased. But increasing the test length also decreased the standard errors of ability estimates because tests became more precise[8]. On the other hand, in the second part of the Research Question 2, increasing the exposure control parameter decreased the overall IPUI values and increased the SEs of the ability estimates. In both of these sets of simulations since only one specification was changed at a time, it was easy to observe the effects of these changes on IPUI and other CAT outcomes. But, it is hard to predict the effect of changing more than one specification on the outcomes of a CAT and IPUI. For instance, if the length of the test and the exposure control parameter have increased at the same time, it would be hard to predict the potential changes in the outcomes. The IPUI values would definitely decrease, but it is very hard to predict the changes in SE. Therefore, it would be valuable to perform multi factor designs where the main effects and the interactions between different CAT specifications on CAT outcomes and IPUI can be observed. As a future expansion of this study, this will be valuable.

6.4.2 A Recommended Value for IPUI

IPUI values are bounded between 0 and 1. This is very useful from the perspective of comparing different item pools or comparing the same item pool for different test specifications or examinee populations. But from a practical point of view it is desirable to have a recommended value for IPUI. If IPUI goes below a specific value, this will signal the test

[7]These constraints were investigated in Section 2.3.3 on page 12

[8]Howeveer, the efficiency of the tests decreased.

developer that the item pool is not sufficient for an examinee or a group of examinees. One of the aims of this study was finding a specific number so that the test developers could use as a flag for an inadequate item pool.

For very basic CAT designs a recommended value for IPUI might be tenable. Section 5.1.1.2 on page 69 explained the relationship between reliability and the standard error. Assuming that the examinee population has a standard normal distribution, a standard error value of 0.32 corresponds to a reliability coefficient of 0.9. Table 5.1 on page 55 indicates that, when the discrepancy between the examinee ability distribution and the item difficulty distribution was 1.5 (or -1.5), the mean standard error of ability estimates were approximately equal to 0.32. The mean IPUI value for these discrepancy conditions were 0.9 and 0.91 respectively.

In the second part or Research Question 1, when the item pool size was 60, the mean SE was approximately 0.30 (see Figure 5.2 on page 56). The mean IPUI value for an item pool size of 60 items was 0.9028 (see Table 5.2 on page 67). These two findings provides some evidence for a recommended value for IPUI. For these particular test designs and examinee populations there was a correspondence between a reliability of 0.9 and an IPUI value of 0.9. But the results of Research Question 2 -where the effects of test specifications on IPUI were investigated- did not confirmed this kind of relationship between the mean SE and mean IPUI[9].

Additionally, attaching a direct link between IPUI and another outcome of a CAT test such as SE will obviate the use of IPUI. IPUI captures a unique aspect of the item pool, whether it is adequate or not. Other outcomes of the CAT are capturing some other important aspects of the test, but not necessarily the adequacy of the item pool. Also, there may not be a direct link between IPUI and the other outcome of the CAT. For example, Section 3.4 showed that IPUI and SE captures different aspects of the CAT. Results of the first part of the Research Question 2 showed that a decrease in the adequacy of an item pool did not imply a decrease in the quality of the measurement. As test lengths increased, the mean

[9]See Figures 5.14 and 5.22 on page 74 and on page 87.

values of IPUI and SE decreased[10].

The challenge with finding a recommended value is closely related to the definition of an inadequate item pool. Unfortunately there is not a clear definition of an insufficient item pool. An insufficient item pool reveals itself by the outcomes of the test. These can be high standard errors of the ability estimates, the bias of the ability estimates, the violation of the constraints of the test or failure to satisfy the test specifications. There are no universally accepted benchmarks for any of these outcomes. Obviously, tests that meet all of the test specifications and having low standard errors of the ability estimates and biases are desirable. But how low is good enough? If there was an accepted threshold between a good test and a bad test, it could be possible to find a specific IPUI value for that threshold.

This challenge applies to other indices of the test quality as well. If we ask "What is a good value for the test reliability?", the answer of a psychometrician would be "It depends on the context.". Here is an excerpt from Nunnally and Bernstein (1994) on the standards of reliability:

> A satisfactory level of reliability depends on how a measure is being used. In the early stages of predictive or construct validation research, time and energy can be saved using instruments that have only modest reliability, e.g., .70. [...] In contrast to the standards used to compare groups, a reliability of .80 may not be nearly high enough in making decisions about individuals. Group research is often concerned with the size of correlations and with mean differences among experimental treatments, for which a reliability of .80 is adequate. [...] If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard. (pp. 264-265)

As the authors indicated, the recommended value for reliability depends on the context

[10]See Figure 5.14 on page 74.

where scores will be used. Similarly, the recommended values for IPUI should depend on the context where the item pool will be used. A high stakes adaptive test might require an IPUI value of .99 for each examinee. For example, the operational pools of the NCLEX-RN exam were adequate close to the cut score. The IPUI values of the examinees were larger than 0.99 close to the cut score (Figure 5.44). On the other hand, if the purpose of the adaptive test is simply to obtain a reliable group summary of examinees, then a lower value of IPUI might be acceptable (Kruyen, Emons, & Sijtsma, 2012).

In addition, the recommended values for reliability mentioned earlier was possibly not agreed upon right away among researchers when Cronbach wrote his highly cited paper in 1951 (Cronbach, 1951)[11]. Instead, years of use of the internal reliability coefficient in different contexts established these recommended values among researchers. Similarly, it is expected that as the use of IPUI becomes prevalent among the practitioners and it is used in different contexts, the properties of IPUI and it's interactions with other test specifications will be explored more. This will potentially lead to a recommended value for IPUI.

6.4.3 Detection of the Redundant Items in the Item Pool

There is a delicate balance between a satisfactory item pool and an item pool which has more than enough items. Both of them serve well for the purposes of a test. But larger item pools have their own disadvantages as discussed in Section 2.3.5.1. Test developers want their item pools to satisfy the test purposes sufficiently. However, due to the cost of developing items, they don't want redundant and underused items in the item pools. Building an item pool that contains just enough number of items is not easy .

IPUI cannot distinguish between a lavish item pool which has more than enough items and an item pool that is satisfactory enough and does not have redundant items. For both of these item pools, IPUI will be 1. If an item pool had an IPUI value of 1, adding more items to this item pool would not increase the value of IPUI further. In practice, an IPUI

[11]Here, it is not suggested that Cronbach is the inventor of reliability.

value of 1 almost never happens. Unless the test developer adds items with exactly the same item parameters, mean IPUI values will always increase. See Table 5.2 on page 67 on how an increase in the size of the item pool increased the IPUI slightly for large item pool sizes.

6.4.4 The Purpose of the Test and the Definition of the Optimum Item

IPUI initially developed for the adaptive tests that are designed to measure every examinee as precisely as possible regardless of the ability of the examinees. This goal is the purpose of most of the achievement tests. But not all tests are designed around this purpose.

Licensure tests, for example, primarily interested in whether an examinee is above a cut score or below it. The precision of the ability estimate of an examinee who is far away from the cut score is not crucial as long as the decision regarding the passing status of this examinee is clear. As a result, if an adaptive test does not give the most appropriate item for this examinee, this is tolerable. Theoretically, the best item pool for the purpose of a licensure test is an item pool including sufficient number of items that have item difficulties equal to the cut score. For the purposes of the test, this item pool is perfect, but for the precision of the ability estimates it is far from perfect. Many other examples can be given for tests in which the high precision of the estimates for all examinees is not the primary purpose for the test.

IPUI in Equation (3.5) quantifies the quality of an item pool as if the primary purpose of the adaptive test is the precision of the ability estimates. In the CAT literature, the optimum item defined according to this purpose: "... an item is considered to have optimum statistical properties if it is most informative at an examinees current maximum-likelihood estimate of ability" (Eignor et al., 1993, p. 10). For the general logic of the adaptive test this makes sense. This is why Flaugher (2000) listed a rectangular distribution of item difficulty as a characteristic of a satisfactory item pool. A rectangular distribution of item difficulty enables a CAT procedure to provide each examinee an appropriate item.

In the future, the formulation of IPUI can be generalized to include various purposes of

the adaptive tests. The denominator can be modified so that the optimum item is defined in accordance with the test purpose. For the numerator, the information should be calculated in respect to this optimum item.

6.5 Future Research Directions

6.5.1 A General Framework for IPUI

As discussed in the previous section, the definition of the optimum item might be different for tests with different purposes. IPUI has a limited definition of an optimum item. Future research can investigate a generalized framework for IPUI which encompasses different definitions of the optimum item.

In a generalized framework, the definition of the optimum item does not have to be an item that has the maximum information at examinee's intermediate ability estimate. Instead, the optimum item can be defined in accordance with the purpose of the test. IPUI can quantify the discrepancy between the administered item and the optimum item.

For instance, for a licensure test with one cut score, the optimum item has a difficulty parameter that is equal to the cut score. Such an item increases the decision accuracy, if not the precision of the ability estimates. IPUI can be calculated as the ratio of the information of the administered item at the cut score to the information of the optimum item at the cut score. If there are multiple cut scores in a test (such as basic, proficient and advanced), the definition of the optimum item becomes complicated (Eggen & Straetmans, 2000).

Another definition for optimum item is related to the test anxiety among examinees. One of the criticisms of a CAT is the difficulty of items presented to the examinee. Item selection in a CAT is optimized so that at each step of the CAT, the algorithm administers an item with 50% probability of correct answer at the intermediate ability estimate of the examinee. This continuing challenge throughout the test might cause frustration to some examinees. Eggen and Verschoor (2006) offered a solution to this problem. Instead of selecting items

that have 50% probability of correct response, they offered an item selection algorithm which selects items that have 60% or 70% (or any other desired percentage) probability of correct response. When comparing their algorithm with MFI item selection algorithm, they observed that they did not achieve desired percentages. They attributed the discrepancy between the actual and desired percentages to "a mismatch between the items available in the item bank and the desired percentages in the population" (Eggen & Verschoor, 2006, p. 391). They hypothesized that addition of easier items (in the case where desired percentage was 60%) to item bank would resolve the problem.

At first sight, IPUI might seem to quantify the mismatch they observed. But in fact IPUI would not help in this situation. The main assumption of IPUI is the definition of optimum item. An optimum item is an item which provides the maximum information at an examinees ability level, an item with 50% probability of correct response at the intermediate θ estimate. In the study of Eggen and Verschoor (2006), the definition of optimum item was different. They defined an optimum item as an item that has maximum information at "an ability value at which the examinee with the current ability estimate has a higher or lower success probability" (p. 387). IPUI as defined as in Equation (3.5) could not capture the "mismatch" they desired.

On the other hand, for this particular item selection algorithm there exists a solution for this problem. The formulation of IPUI could be changed to adjust for their definition of optimum item. In Equation (3.5), instead of calculating information at $\hat{\theta}_{k-1}$, the information can be calculated at $\hat{\theta}_{k-1} - \delta_i$, where $\hat{\theta}_{k-1}$ is the intermediate ability estimate before the administration of kth item. δ_i is the shift parameter for the i-th item which is defined for 2PL as $\frac{1}{a_i} \ln(\frac{p}{1-p})$, where p is the desired probability of correct response. For example, if the desired probability is 0.60, the item selection algorithm will select an item that has maximum information at $\hat{\theta}_{k-1} - \delta_i = \hat{\theta}_{k-1} - \frac{1}{a_i} \ln(\frac{0.6}{1-0.6})$.

For 1PL model, the definition of optimum item for an examinee with intermediate ability estimate $\hat{\theta}_{k-1}$ is an item with difficulty parameter equals to $\hat{\theta}_{k-1} - \ln(\frac{p}{1-p})$. When the desired

probability $p = 0.5$, this corresponds to a difficulty parameter equals to the intermediate θ estimate. For 2PL model, the definition of optimum item is more complicated. Since item discrimination parameter does not have an upper bound, the test developer can define a maximum value for a parameter (a_{max}). Accordingly, item discrimination parameter of the optimum item is a_{max} and the item difficulty parameter is $\hat{\theta}_{k-1} - \frac{1}{a_{max}} \ln(\frac{p}{1-p})$. Using the general framework discussed above, the definition of IPUI for the kth administered item i_k will be:

$$IPUI_k = \frac{\mathcal{I}_{i_k}\left[\hat{\theta}_{k-1} - \frac{1}{a_{max}} \ln(\frac{p}{1-p})\right]}{\mathcal{I}_{max}\left[\hat{\theta}_{k-1} - \frac{1}{a_{max}} \ln(\frac{p}{1-p})\right]}$$

This definition can be adjusted for 1PL model by fixing a_{max} to 1. The examples given here can be extended to different optimum item definitions and the IPUI can be used as a more general tool for evaluating the adequacy of item pools for different CAT scenarios.

6.5.2 Weights for IPUI

At individual test level, IPUI is currently giving equal weights at each stage of the adaptive test. In reality, as Chang and Ying (2008) argued, it is desirable for a CAT to provide better items towards the end of the test. The ability estimates at the beginning of the test are prone to more error, so items do not have to match the intermediate ability estimates precisely. Towards to end of the test better items are needed because the ability estimates are more precise. Considering this, the weights of the IPUI might be adjusted to be low at the early stages of the test and high towards the end of the test. In this case, if the item pool is depleted towards the end of the test, where the need for appropriate items is more critical, this weighting scheme will punish the item pool for this deficiency.

6.5.3 IPUI for Other Psychometric Models

IPUI is currently available for only 1PL model. This hampers its use for 2PL and 3PL models which are very common in operational tests. The problem with 2PL and 3PL IRT models

is the parameter values of the optimum item for these models. For an optimum item, the item discrimination parameter (a parameter) should be equal to infinity. The information value of this item also has an infinite value. This makes the denominator of the IPUI infinite. Consequently the value of the IPUI will be undefined.

In practice, such an optimum item is not possible to develop (Reckase, 2010). This limitation can be handled by setting a limit to the item discrimination parameter. Even though the value of this limit will be arbitrary, the historical test data can be used to get this number. For instance, an optimum item for 3PL can be defined as having an item difficulty equal to the intermediate ability estimate, item discrimination equal to 2, and guessing parameter equal to 0.

This approach has some limitations. If an item has an a parameter larger than 2, then the value of IPUI will exceed 1. In addition, the comparison of IPUI values will not be possible if different limits for a parameters are used for different tests.

From the diagnostic point of view, using IPUI only for 1PL makes sense. In reality, when a test developer needs to add an item to the item pool, it is comparatively easier for item writers to write an item that has a targeted item difficulty parameter compared to writing items with a targeted item discrimination parameter (Bejar, 1983). So, diagnostically, if IPUI guides test developers to write item that have some specific difficulty, practically this might be feasible.

In addition to 1PL, 2PL and 3PL models, the use of IPUI can be extended to MIRT models and polytomous IRT models too. A CAT using a MIRT model is more complex compared to the unidimensional CATs (Yao, Pommerich, & Segall, 2014). Consequently, the evaluation of the item pools for multidimensional computerized adaptive test (MCAT) is more difficult. The extension of the IPUI to multidimensional item pools is straightforward because the information function of MIRT is very similar to the information function of unidimensional IRT (Reckase & McKinley, 1991). IPUI can offer an easy way to evaluate the item pools for MCAT.

CATs using polytomous IRT models (Nering & Ostini, 2010) are another possible extension of IPUI. In health sciences, the use of a CAT with polytomous items are common (Amtmann et al., 2010; Haley et al., 2009; Pilkonis et al., 2011). The size of the item pools for CATs with polytomous items are relatively small compared to the item pools used in high stakes tests. IPUI can be helpful for the diagnosis of these small item pools. In addition, due to multiple possible responses for each item, the evaluation of the item pools might be challenging.

6.5.4 Naming of the Index

The correct naming of the index is important because it conveys the message about the possible uses of the index. An index with a misleading name might result in an inappropriate use of the index. The index developed in this study quantifies whether the item pool is adequate for a given set of test specifications and the examinee population. A perfectly adequate item pool might not be adequate for a different set of test specifications or for a different examinee population. The name of the index should convey the dependence of the item pool performance to the test specifications and the examinee population.

The name "item pool utilization index" partially covers this meaning. But in the future, other naming alternatives that conveys the capabilities of this index better should be explored. Some alternative names might be "item pool adequacy index", "quality of utilization of item pool index" and "quality of item pool index". As the use of this index spread among the practitioners and researchers, a consensus on a better name for this index will be reached.

APPENDICES

APPENDIX A

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 1 - PART 1

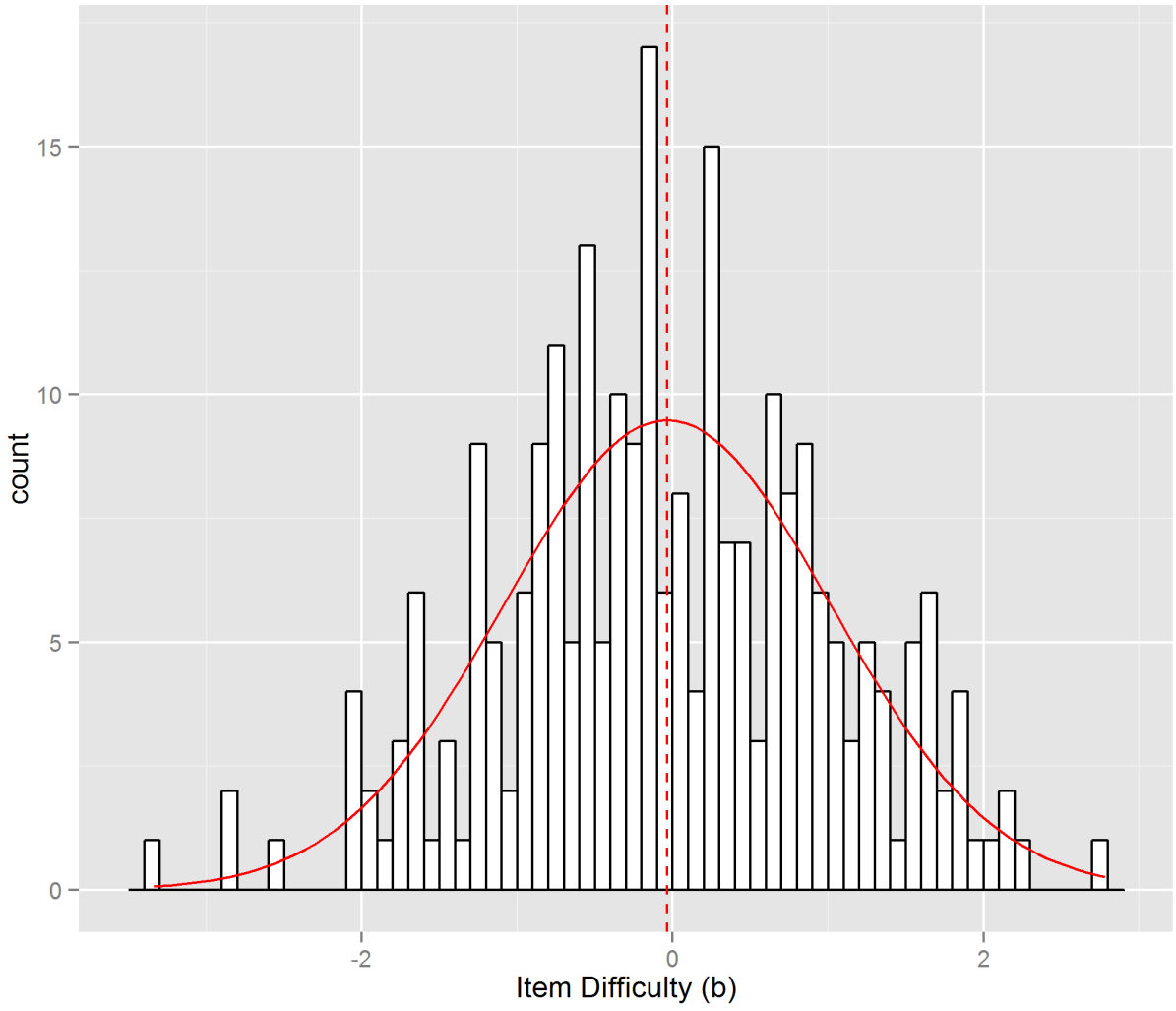

Figure A.1: Item Difficulty Distribution (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

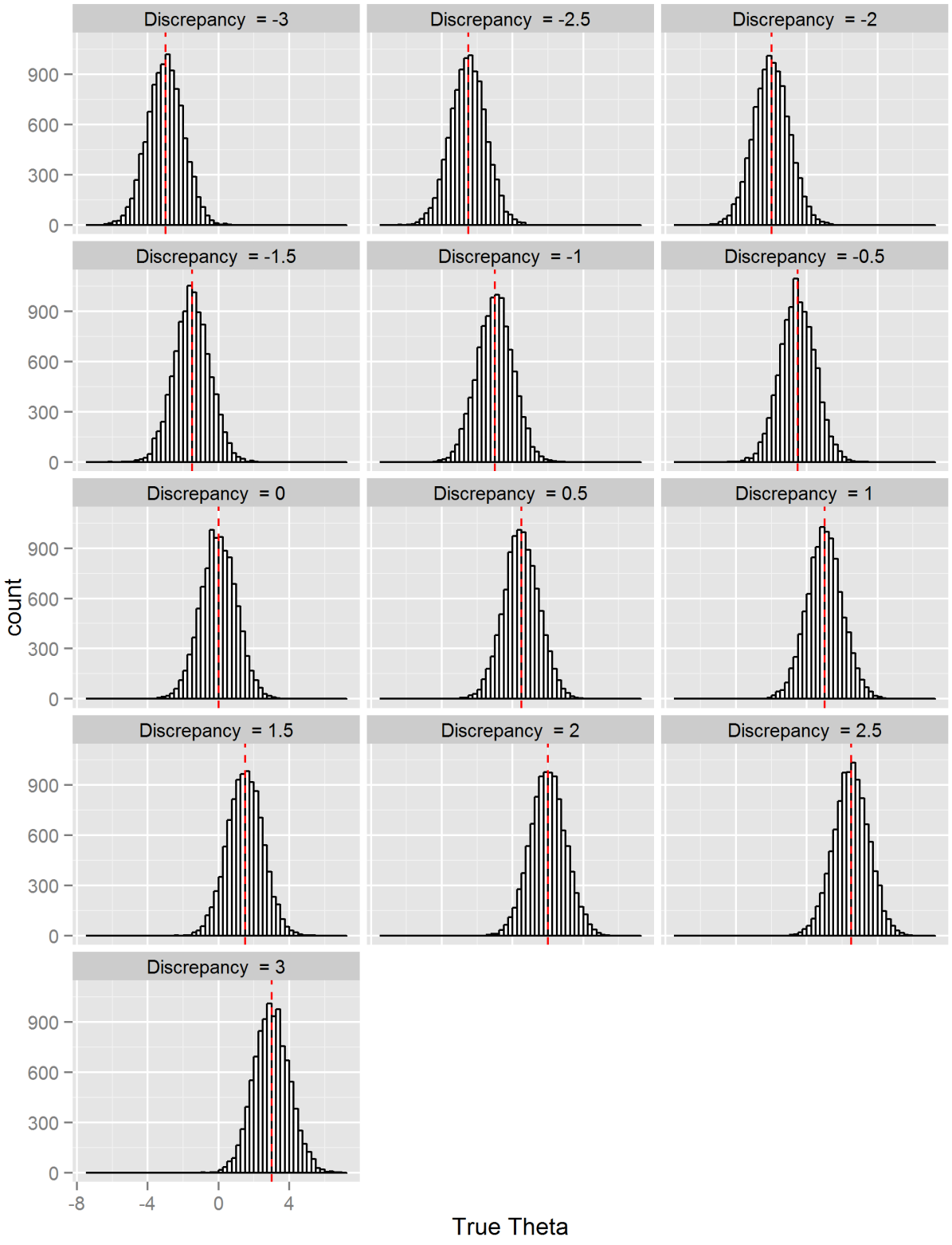

Figure A.2: True θ Distribution (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

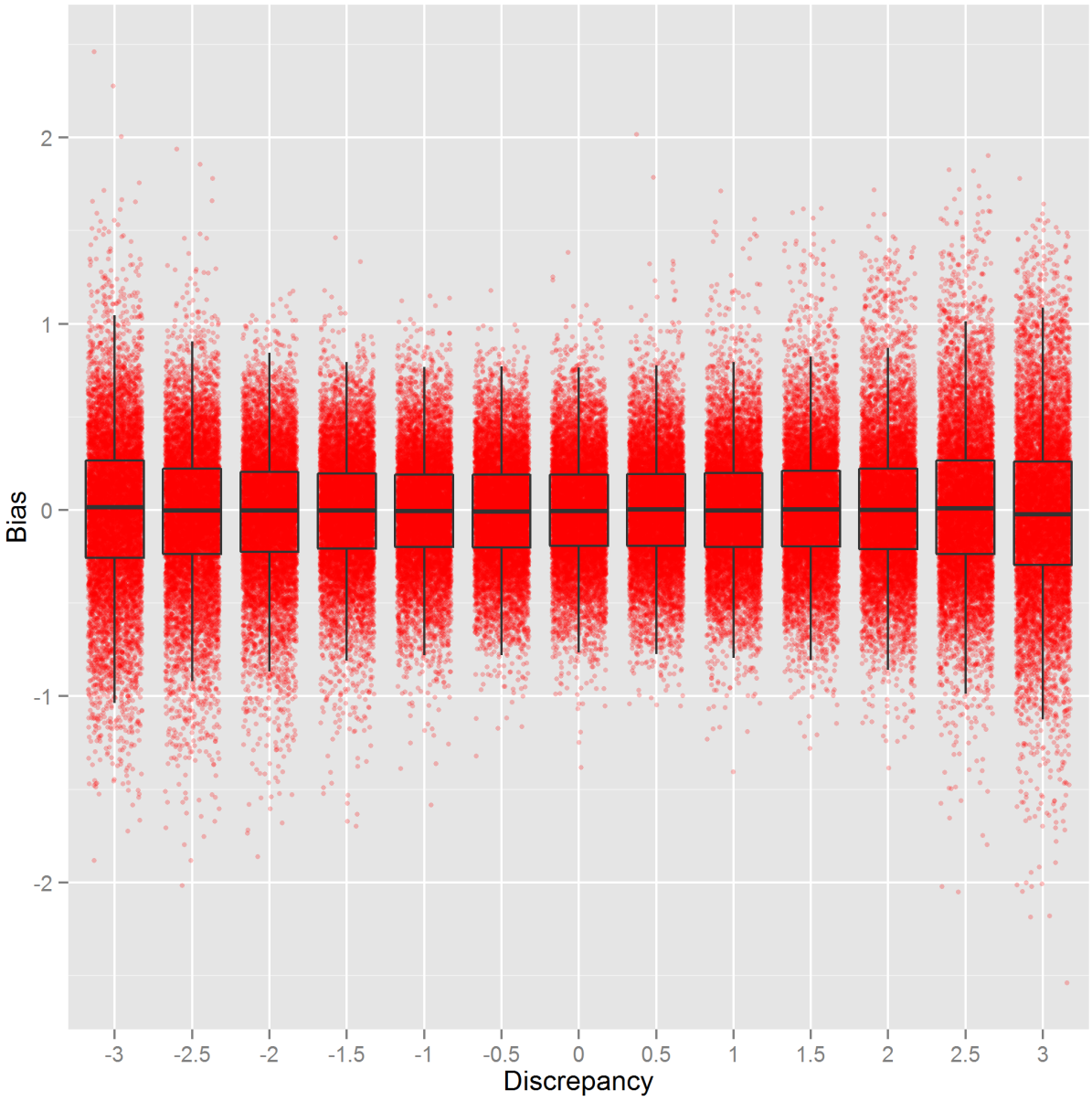

Figure A.3: Distribution of Bias for each Discrepancy Condition (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

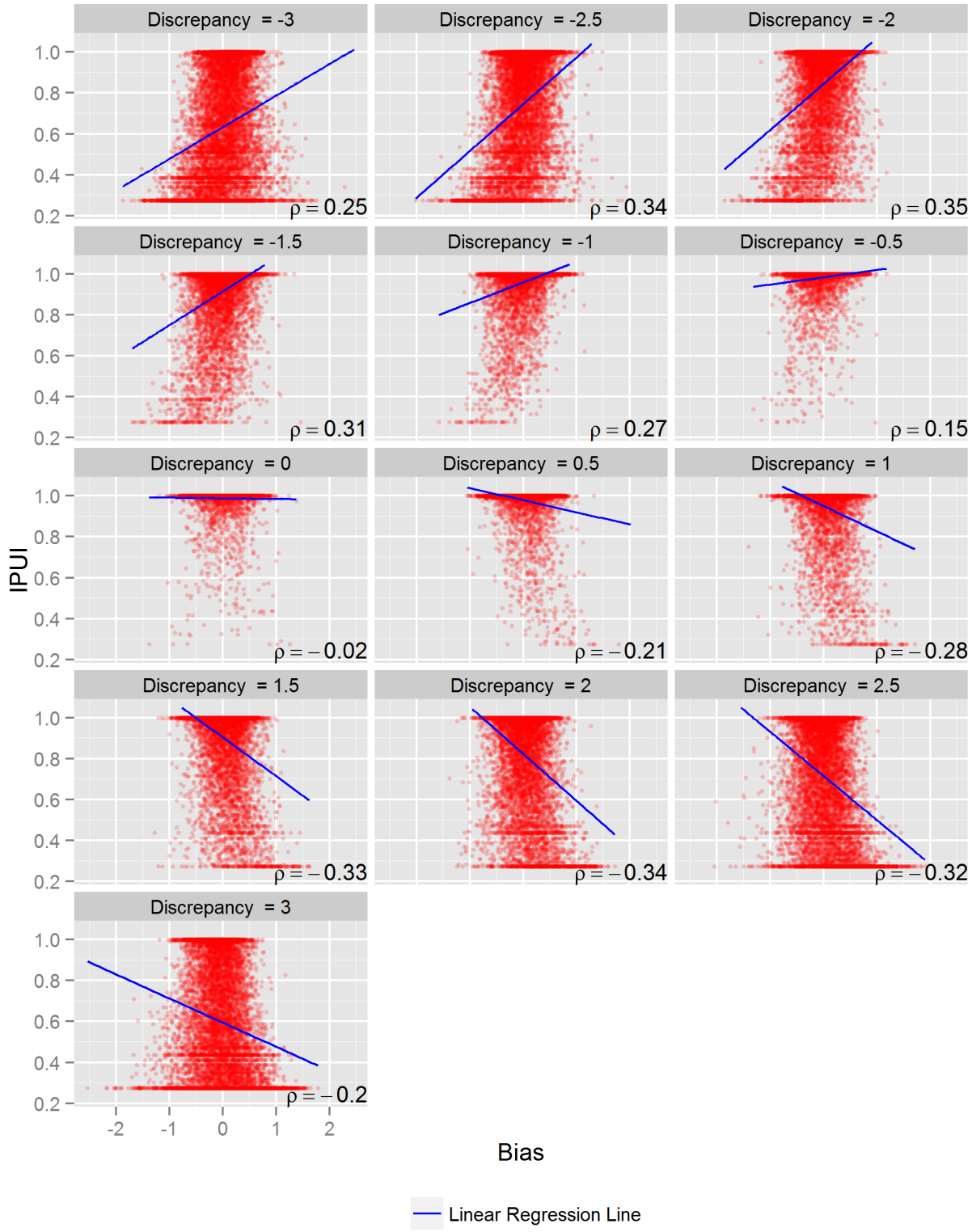

Figure A.4: Relationship between Bias and IPUI for each Discrepancy Condition (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

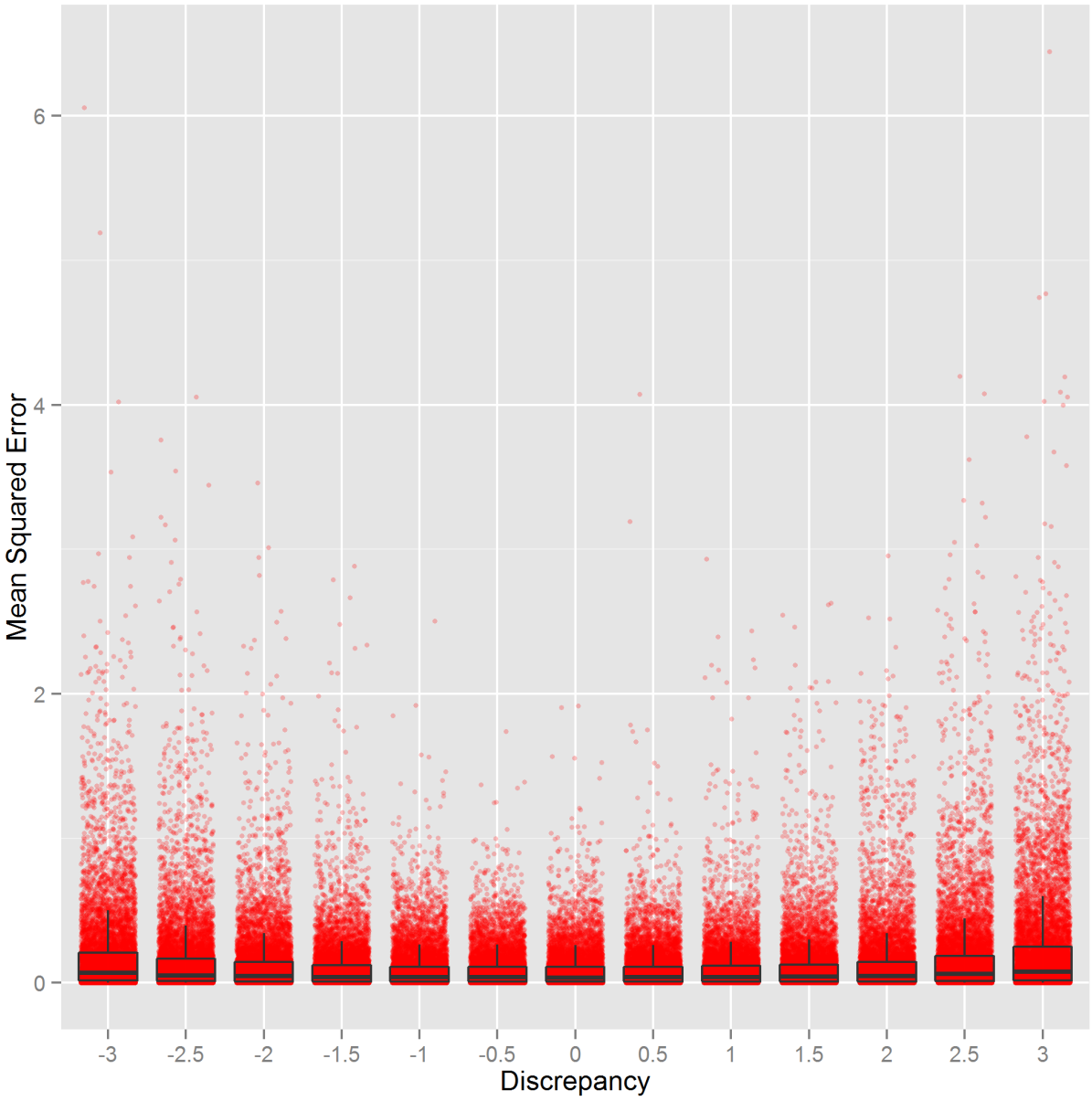

Figure A.5: Distribution of Mean Squared Error for each Discrepancy Condition (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

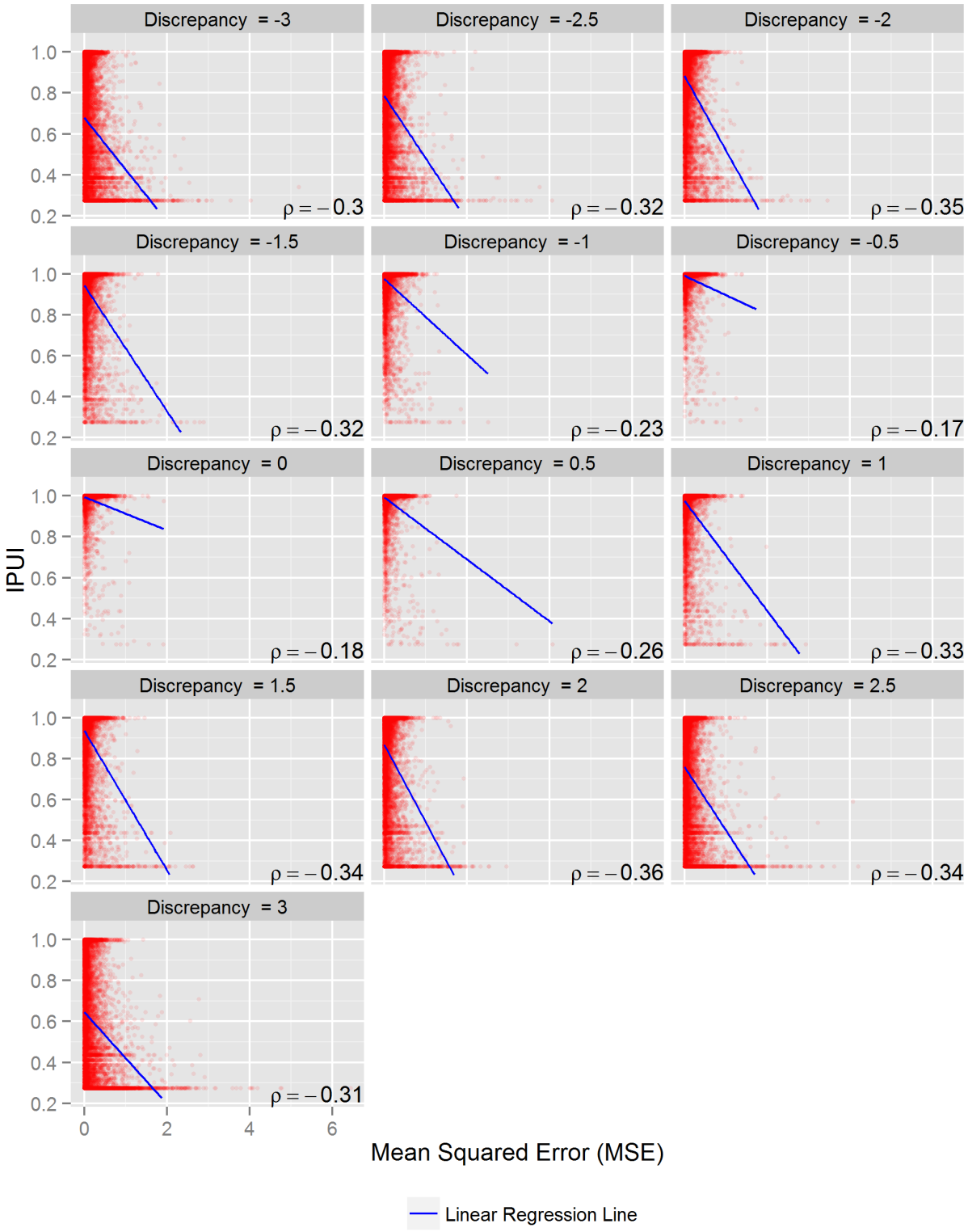

Figure A.6: Relationship between Mean Squared Error and IPUI for each Discrepancy Condition (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

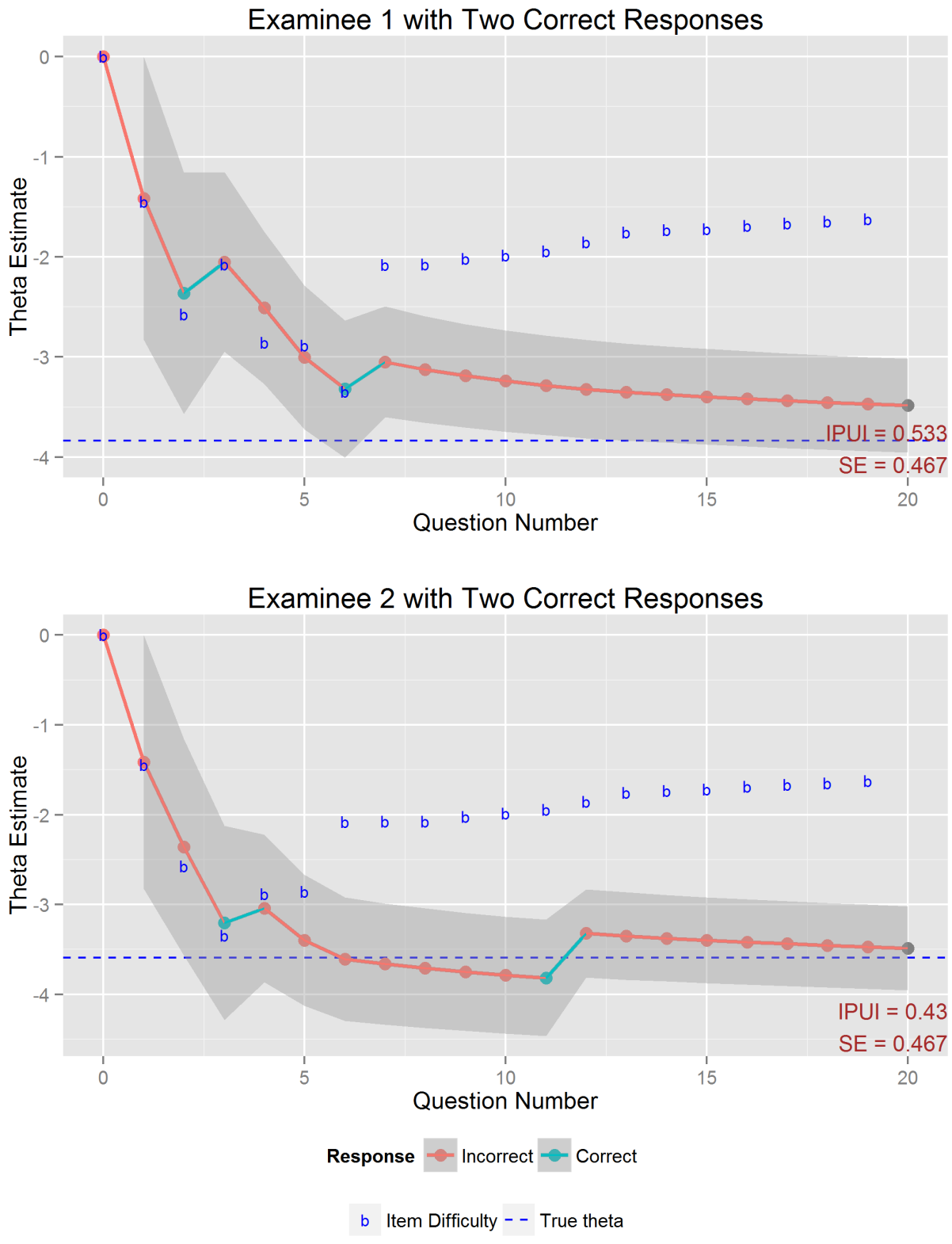

Figure A.7: Two Examinees with Same Standard Errors but Different IPUI Values (Research Question 1 - Discrepancy between Item Pool and Ability Distribution)

APPENDIX B

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 1 - PART 2

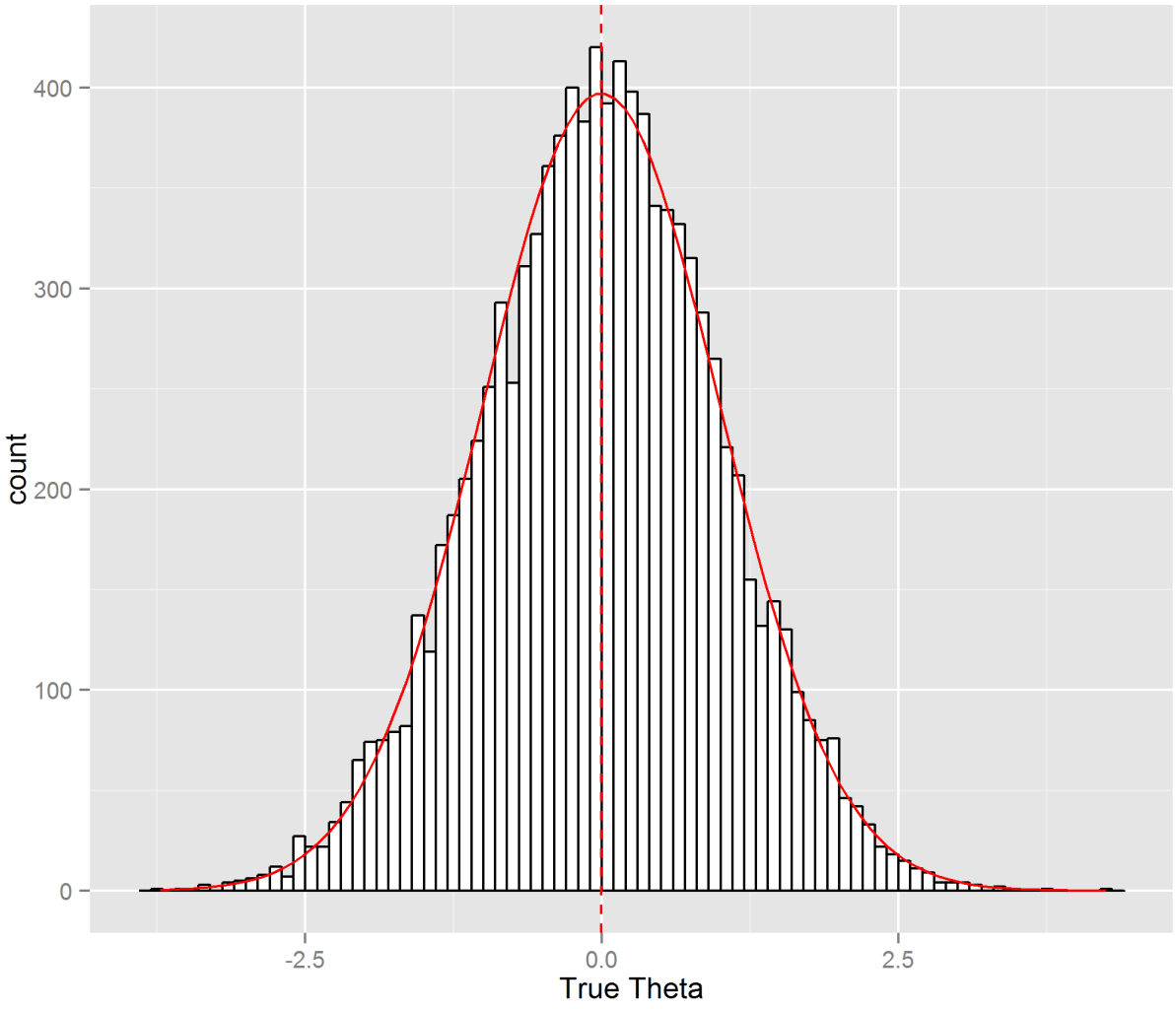

Figure B.1: True θ Distribution (Research Question 1 - Part 2)

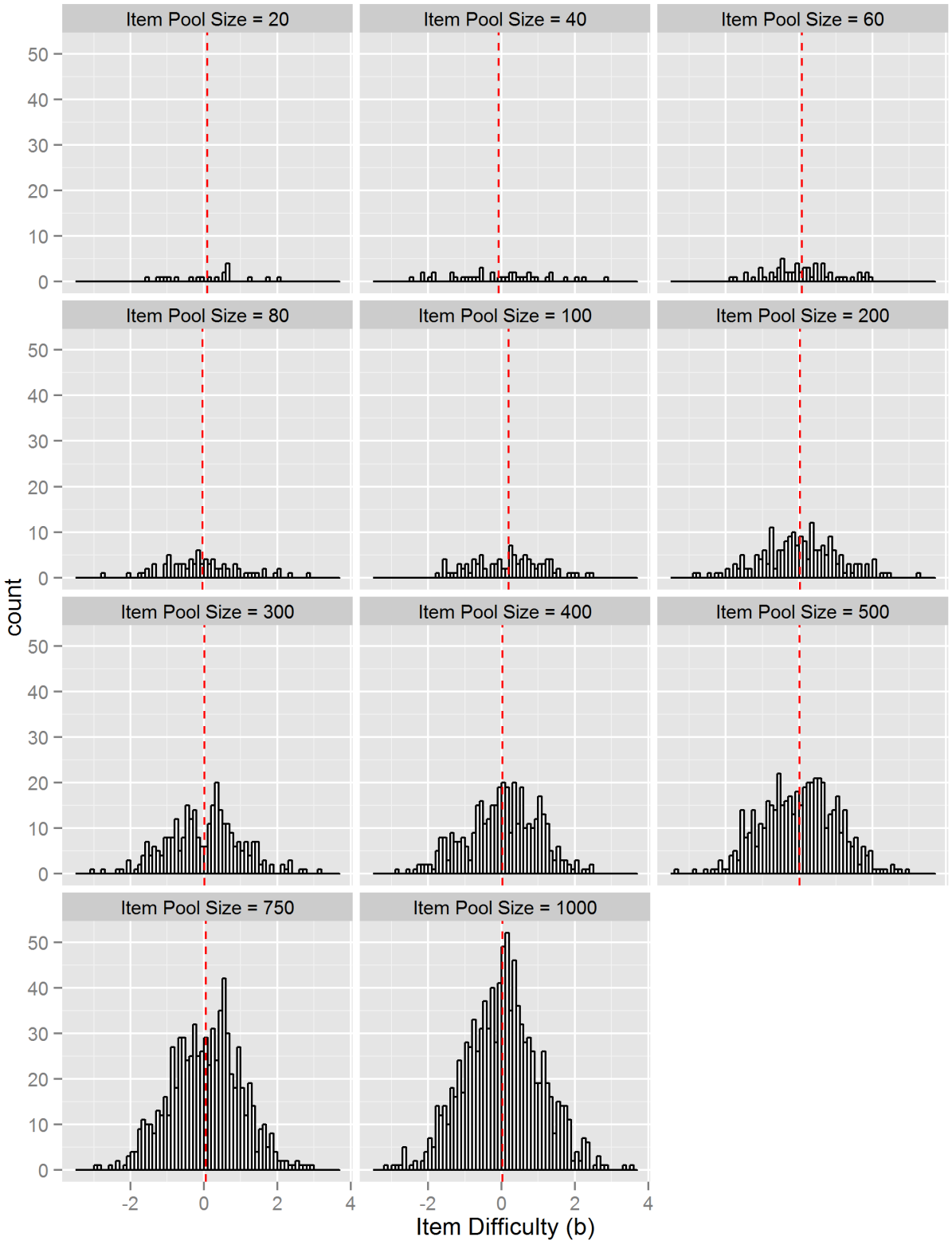

Figure B.2: Item Difficulty Distribution by Item Pool Size Condition for Replication 19 (Research Question 1 - Part 2)

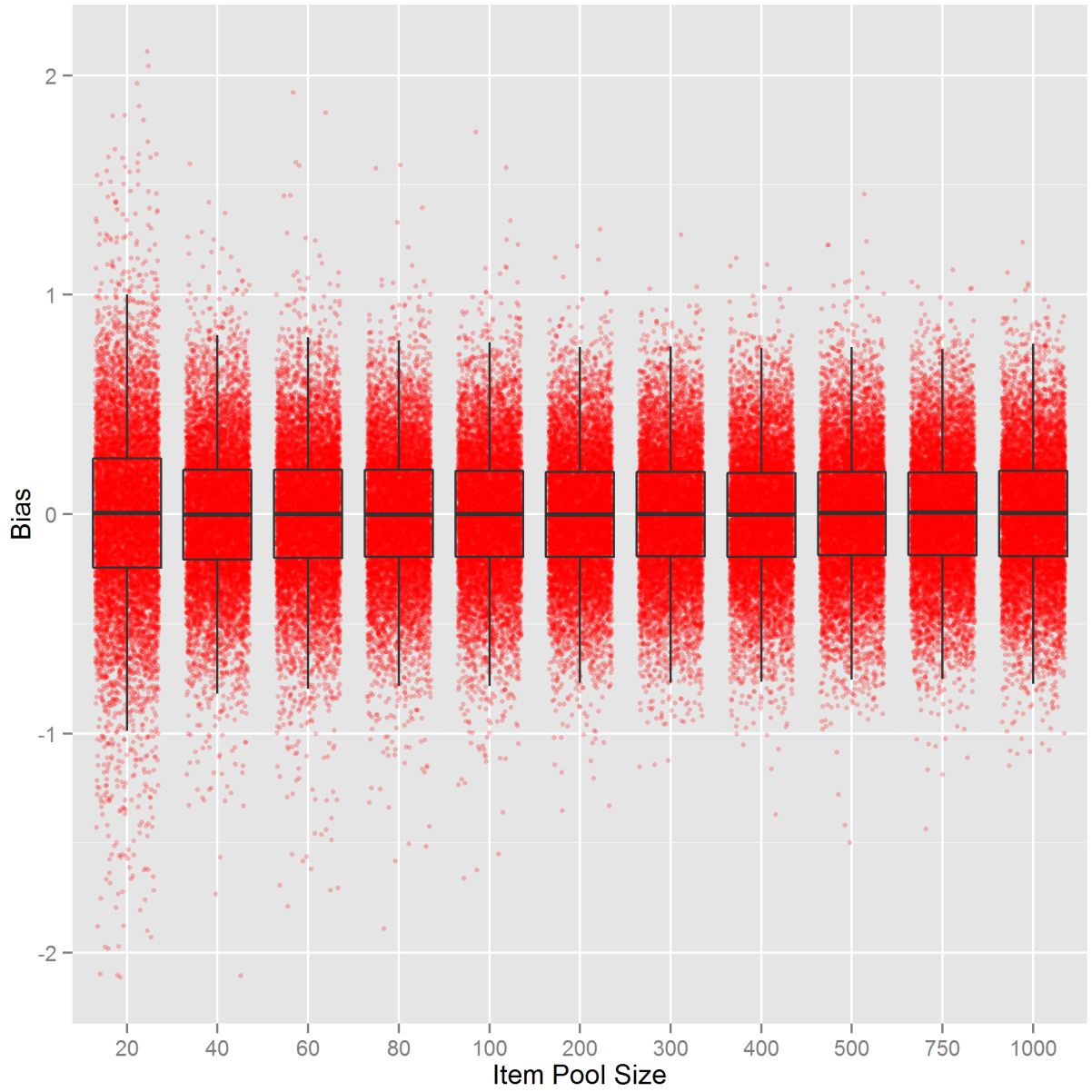

Figure B.3: Bias Distribution by Item Pool Size Condition for Replication 19 (Research Question 1 - Part 2)

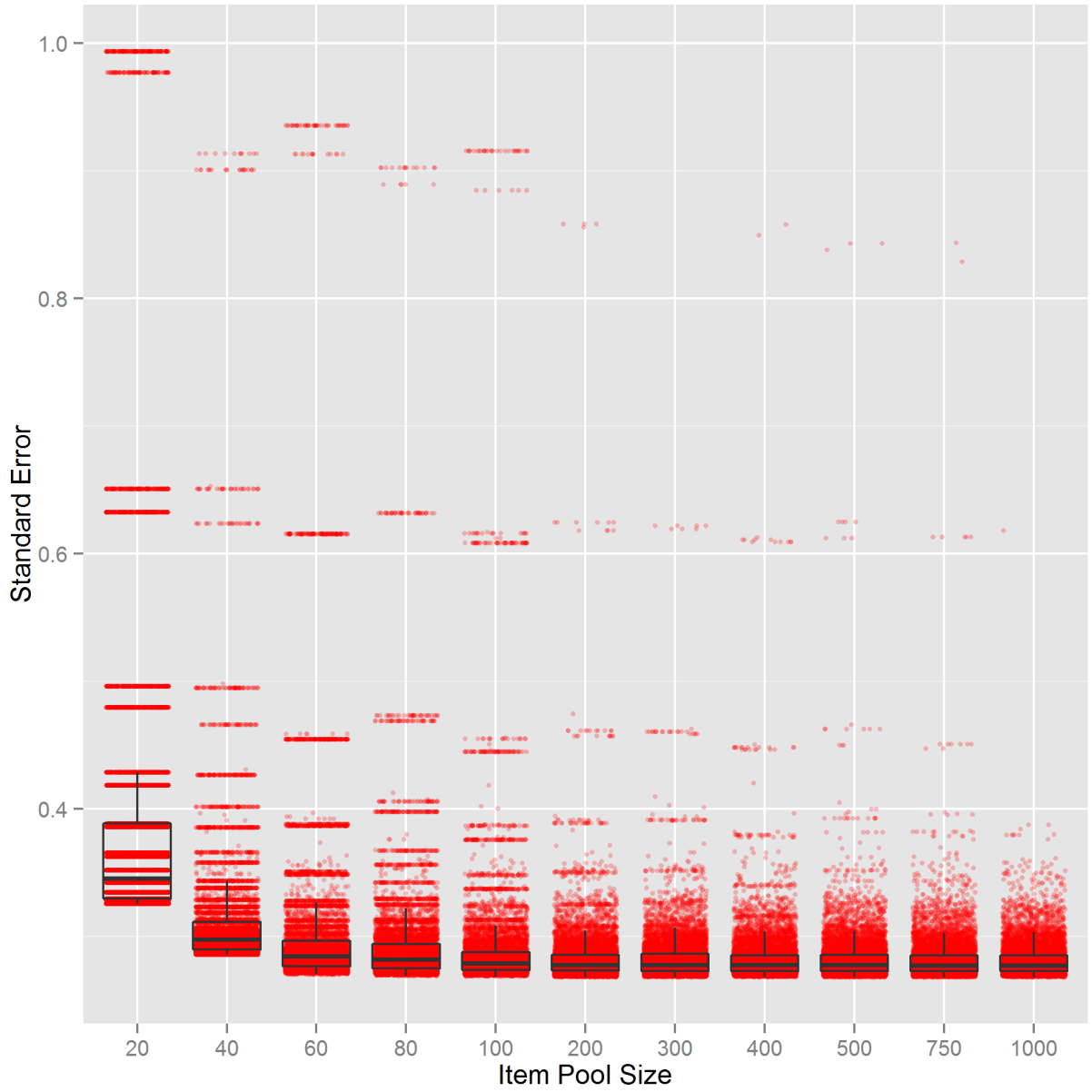

Figure B.4: Standard Error Distribution by Item Pool Size Condition for Replication 19 (Research Question 1 - Part 2)

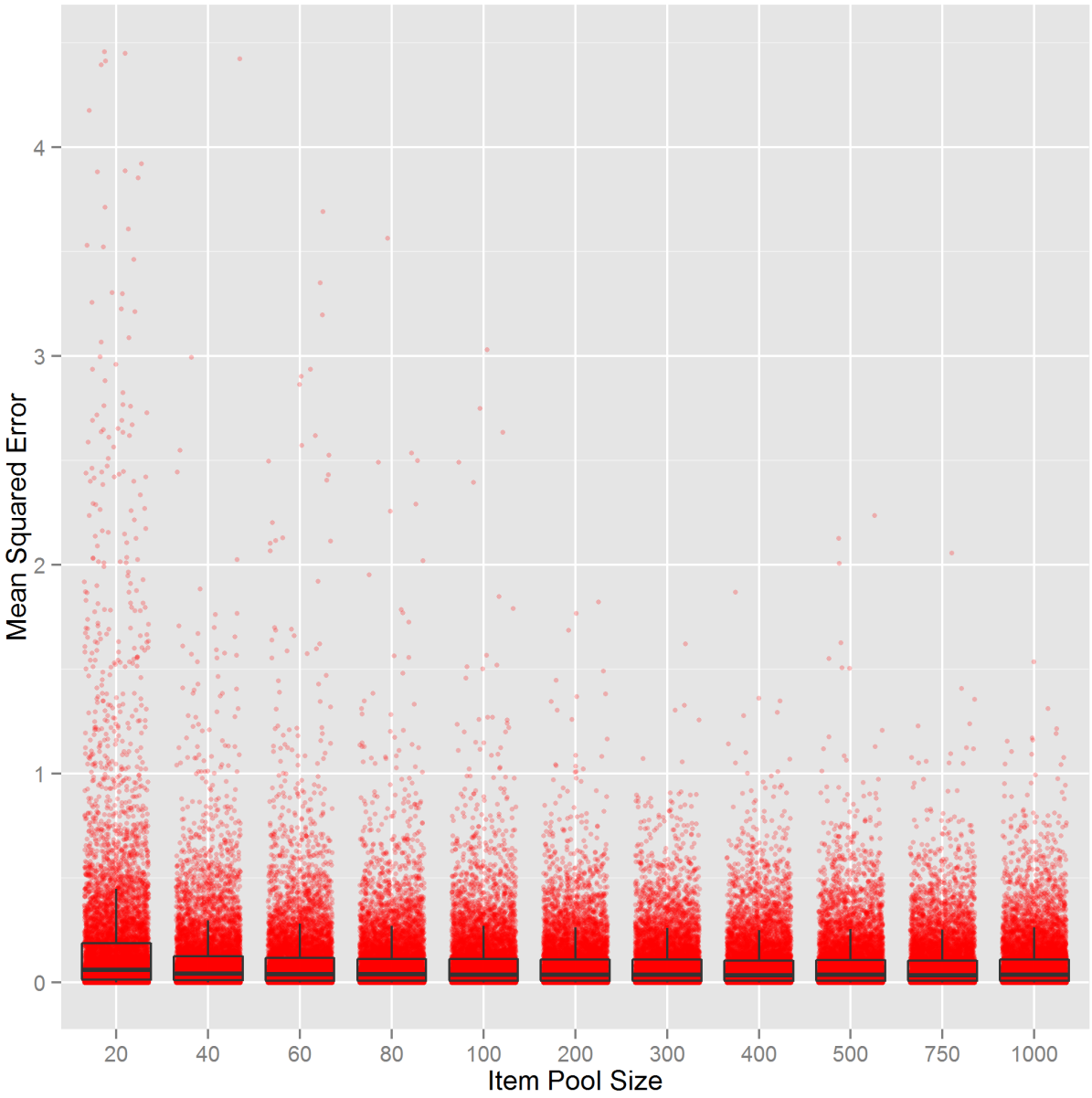

Figure B.5: Mean Squared Error Distribution by Item Pool Size Condition for Replication 19 (Research Question 1 - Part 2)

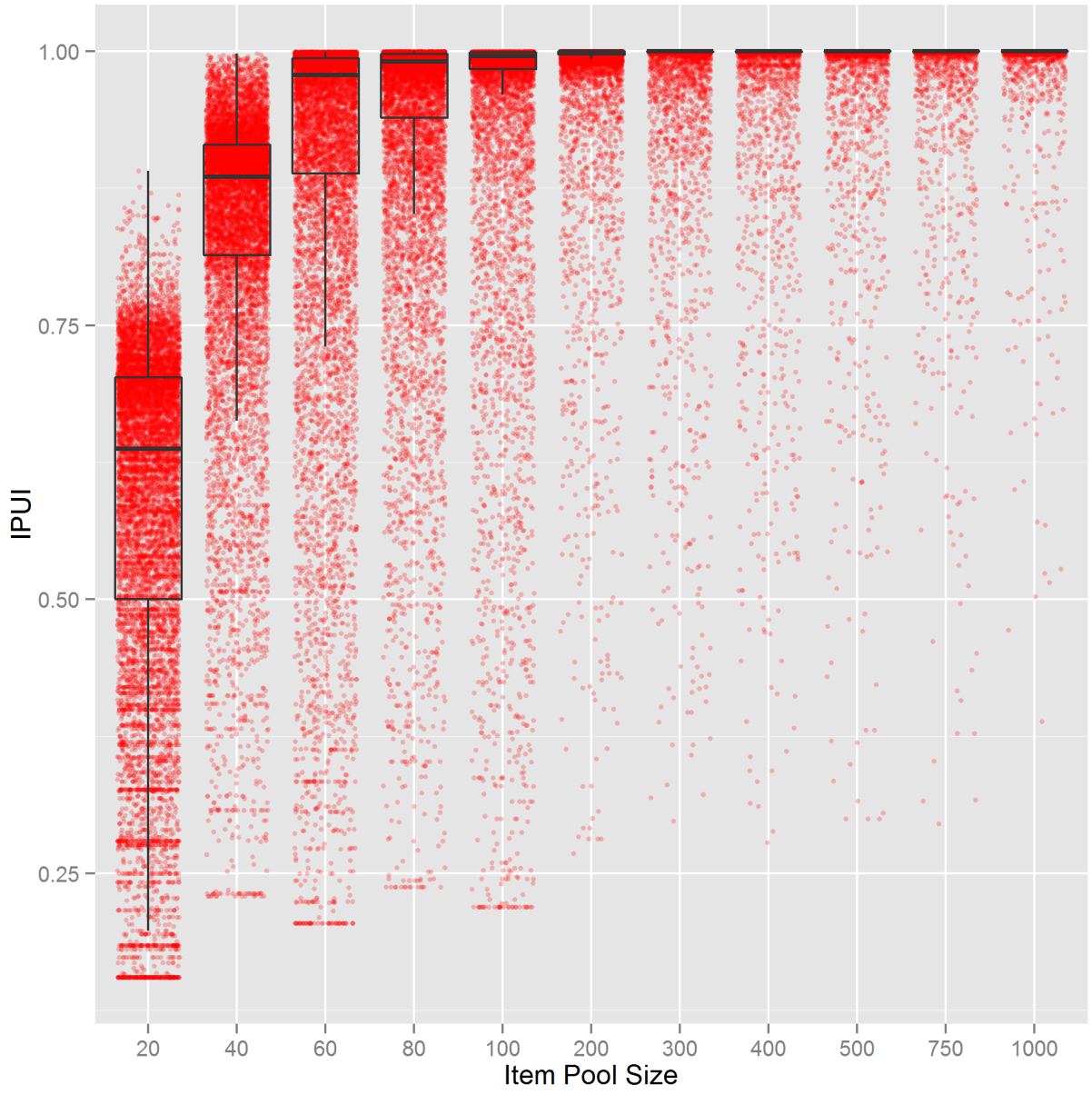

Figure B.6: IPUI Distribution by Item Pool Size Condition for Replication 19 (Research Question 1 - Part 2)

APPENDIX C

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 2 - PART 1

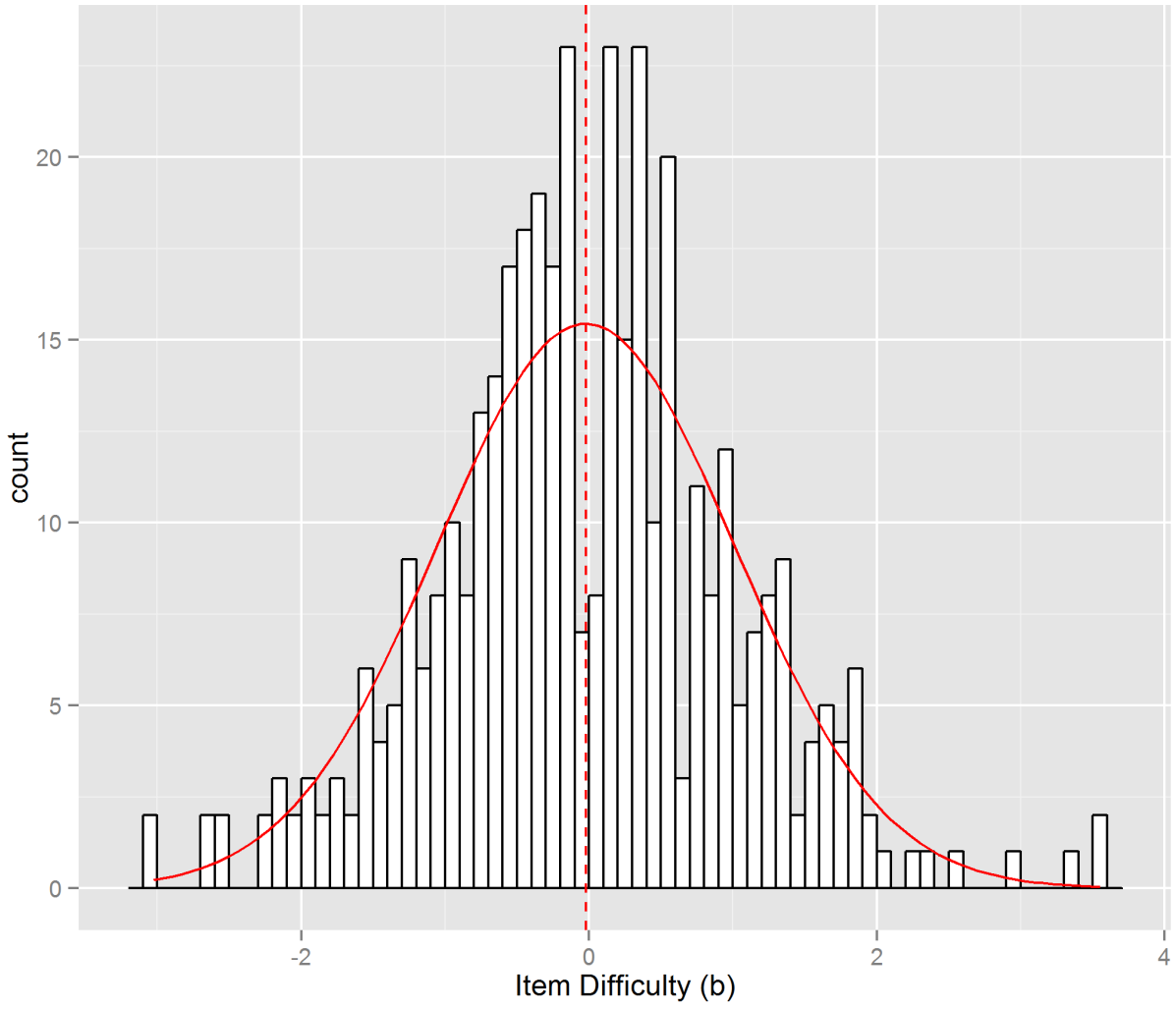

Figure C.1: Item Difficulty Distribution for Research Question 2 - Test Length Conditions

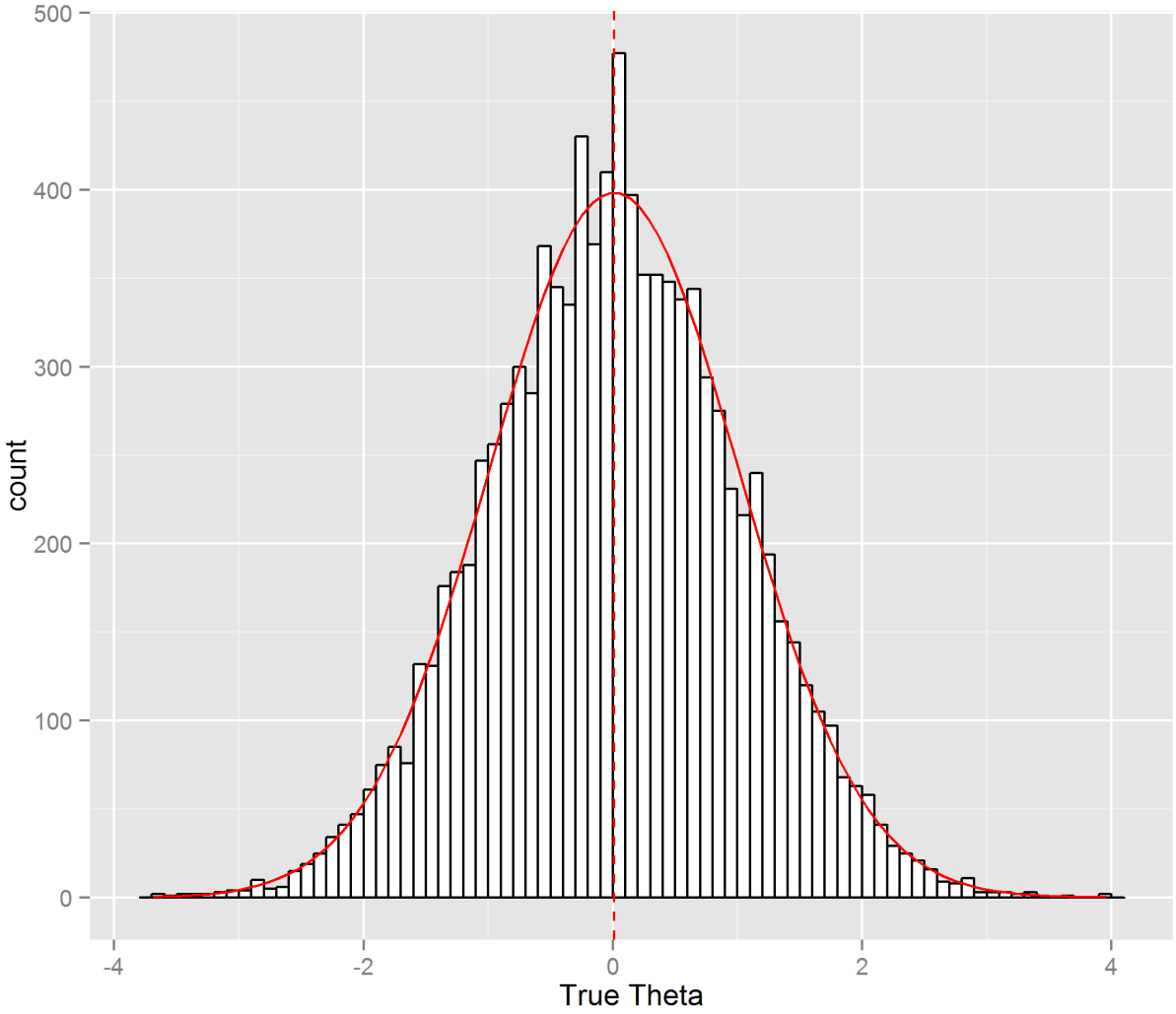

Figure C.2: True θ Distribution for Research Question 2 - Test Length Conditions

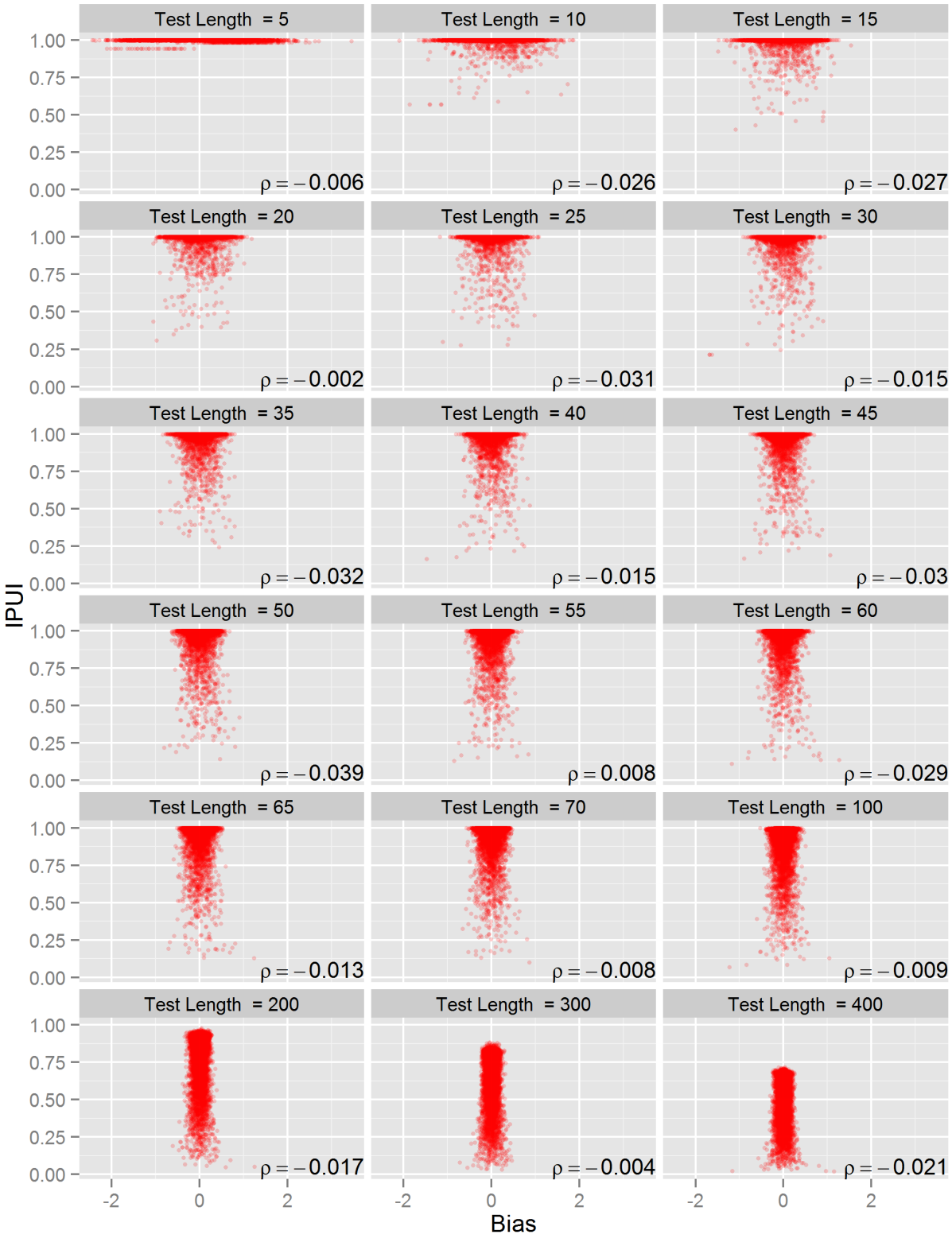

Figure C.3: IPUI and Bias Relationship by Test Length Condition

APPENDIX D

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 2 - PART 2

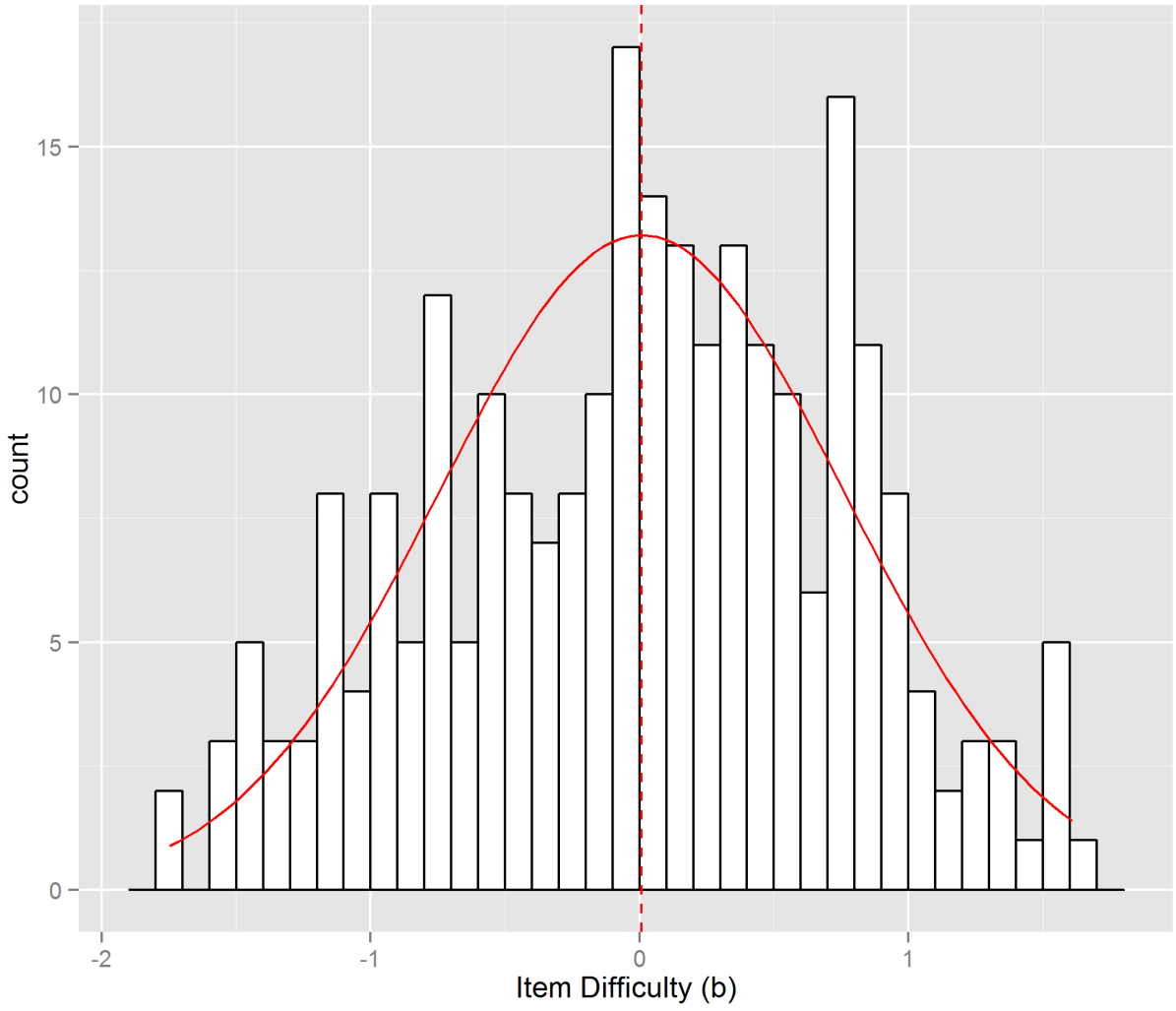

Figure D.1: Item Difficulty Distribution for Research Question 2 - Exposure Control

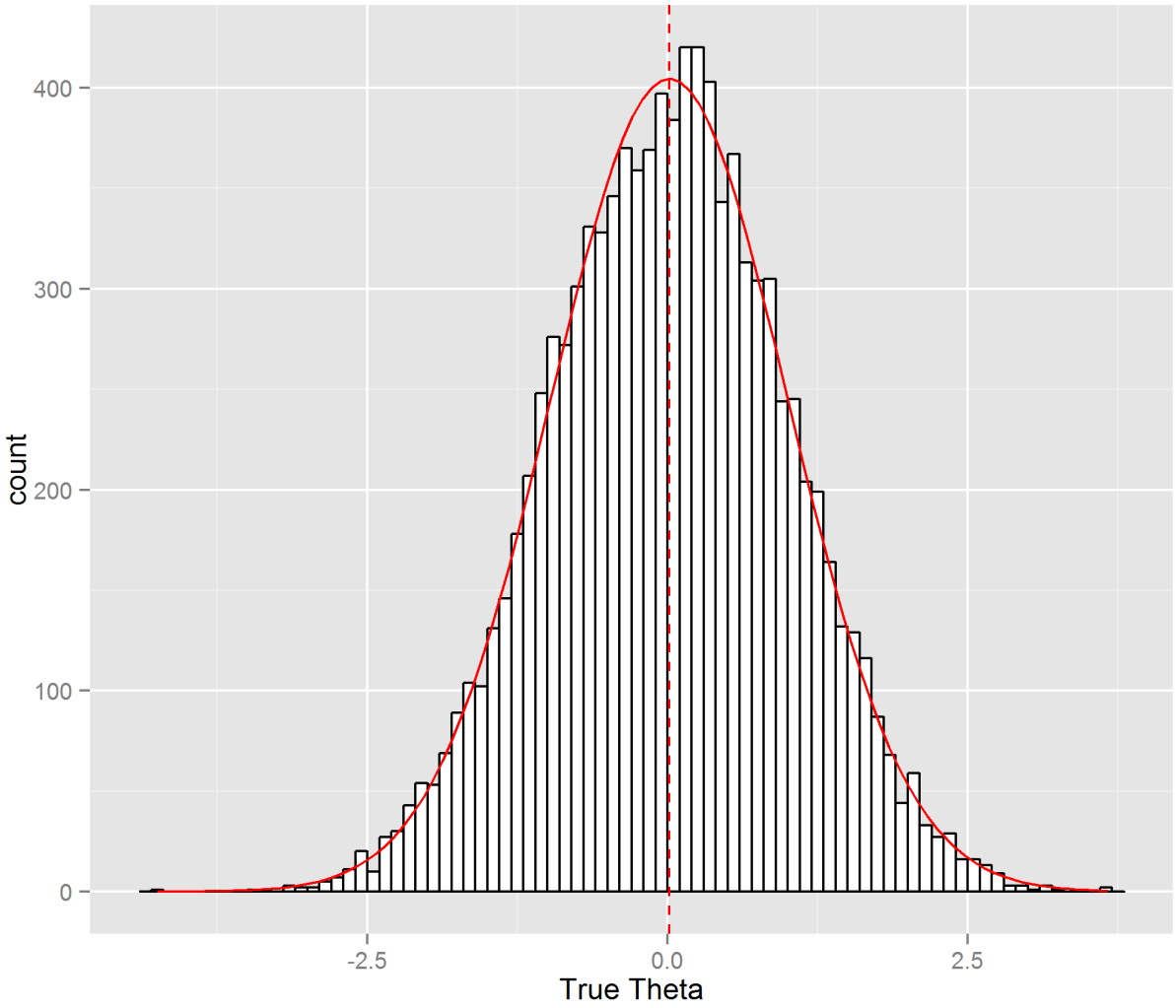

Figure D.2: True θ Distribution for Research Question 2 - Exposure Control

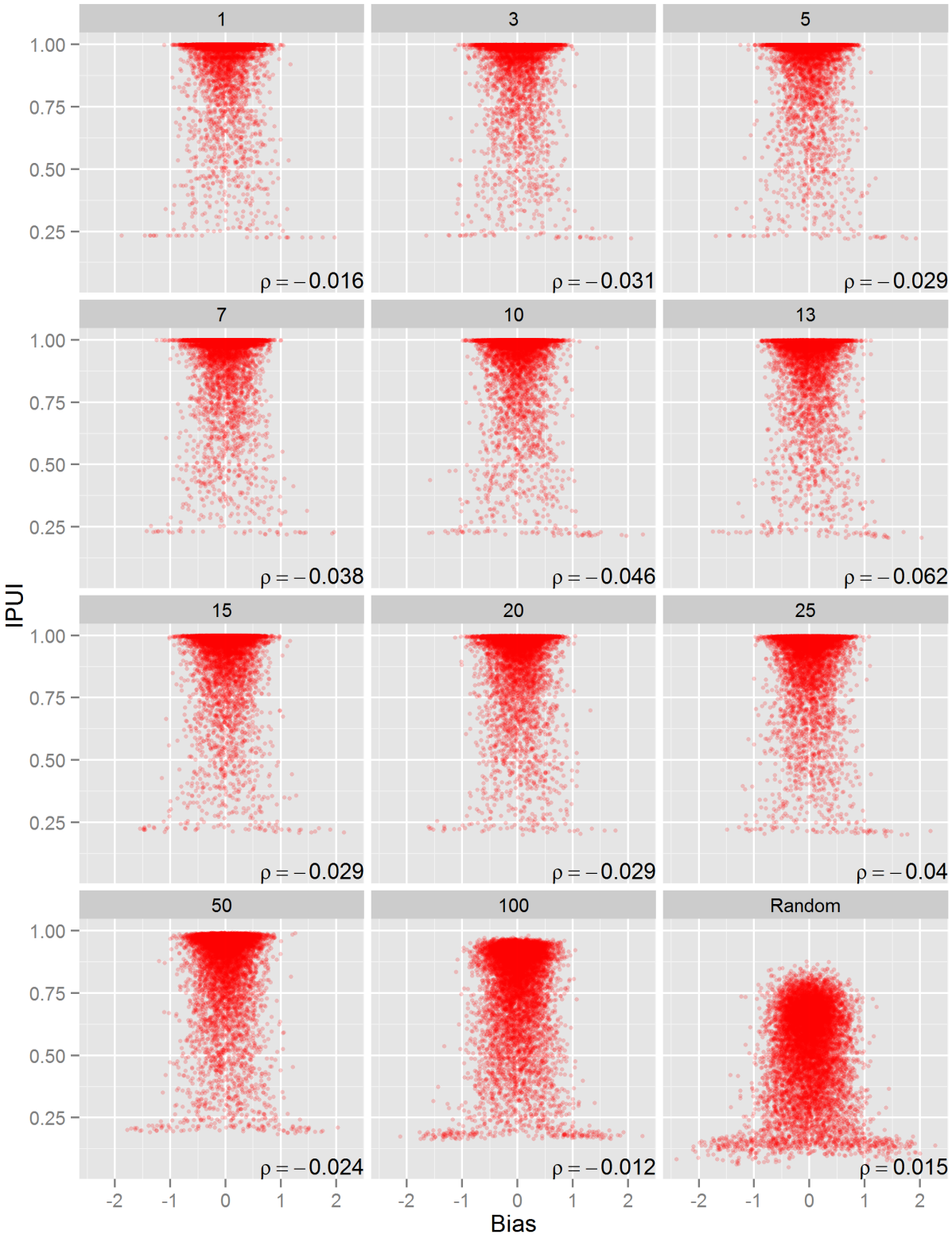

Figure D.3: IPUI and Bias Relationship by Exposure Control Condition

APPENDIX E

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 3

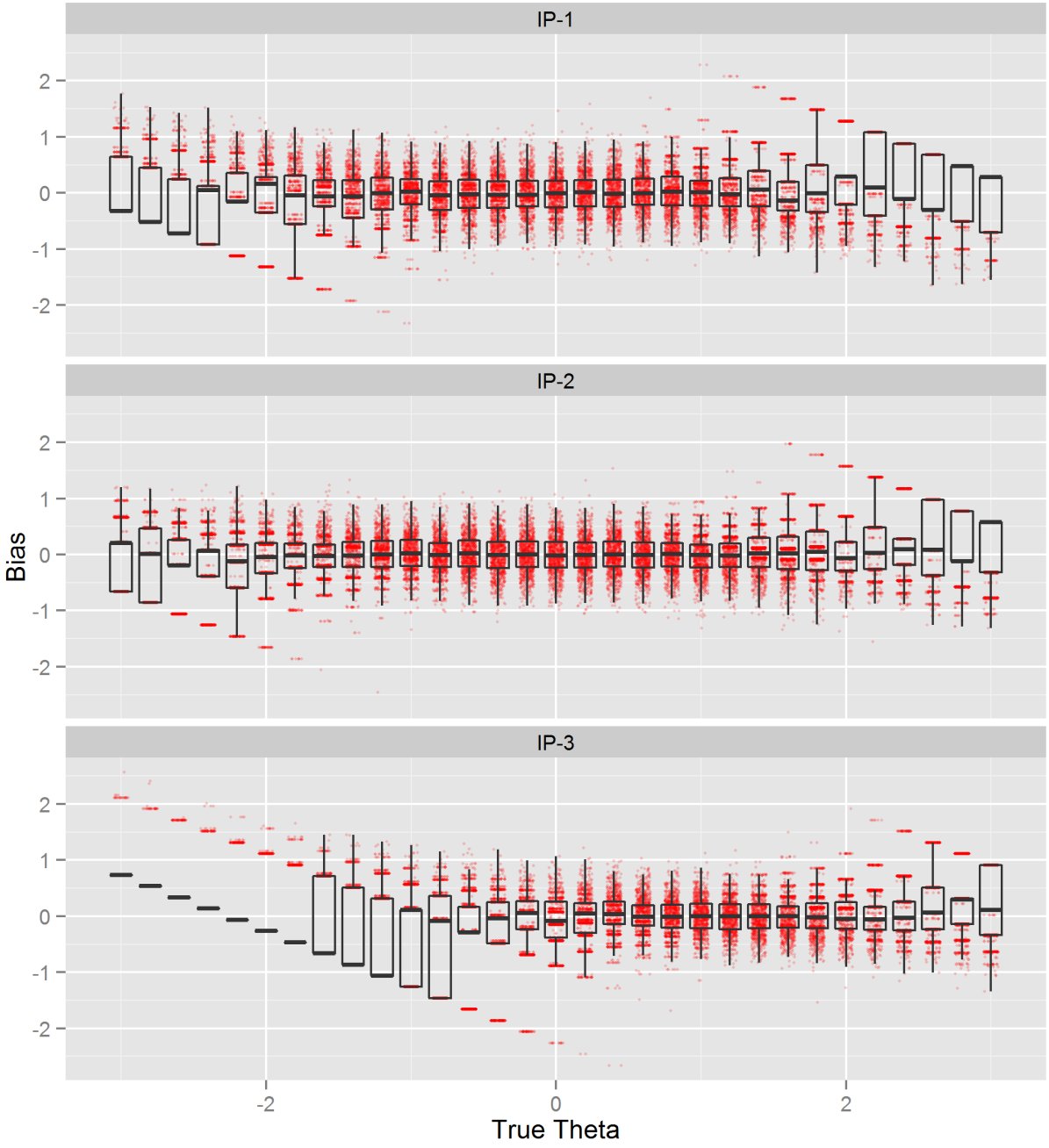

Figure E.1: The Bias Distribution at each True θ Value for each Item Pool Condition

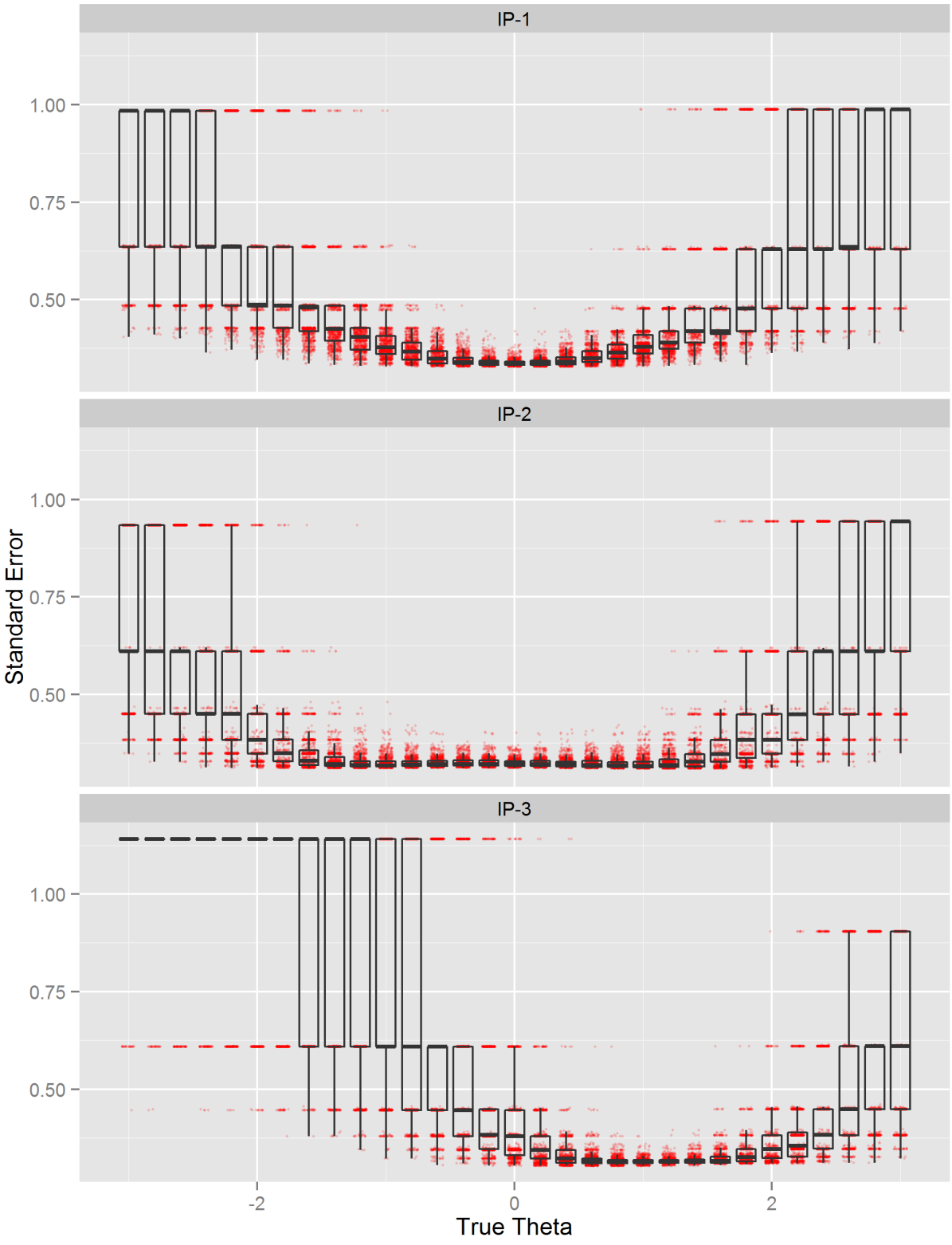

Figure E.2: The Standard Error Distribution at each True θ Value for each Item Pool Condition

APPENDIX F

SUPPLEMENTARY FIGURES FOR IDEAL ITEM POOL CREATION

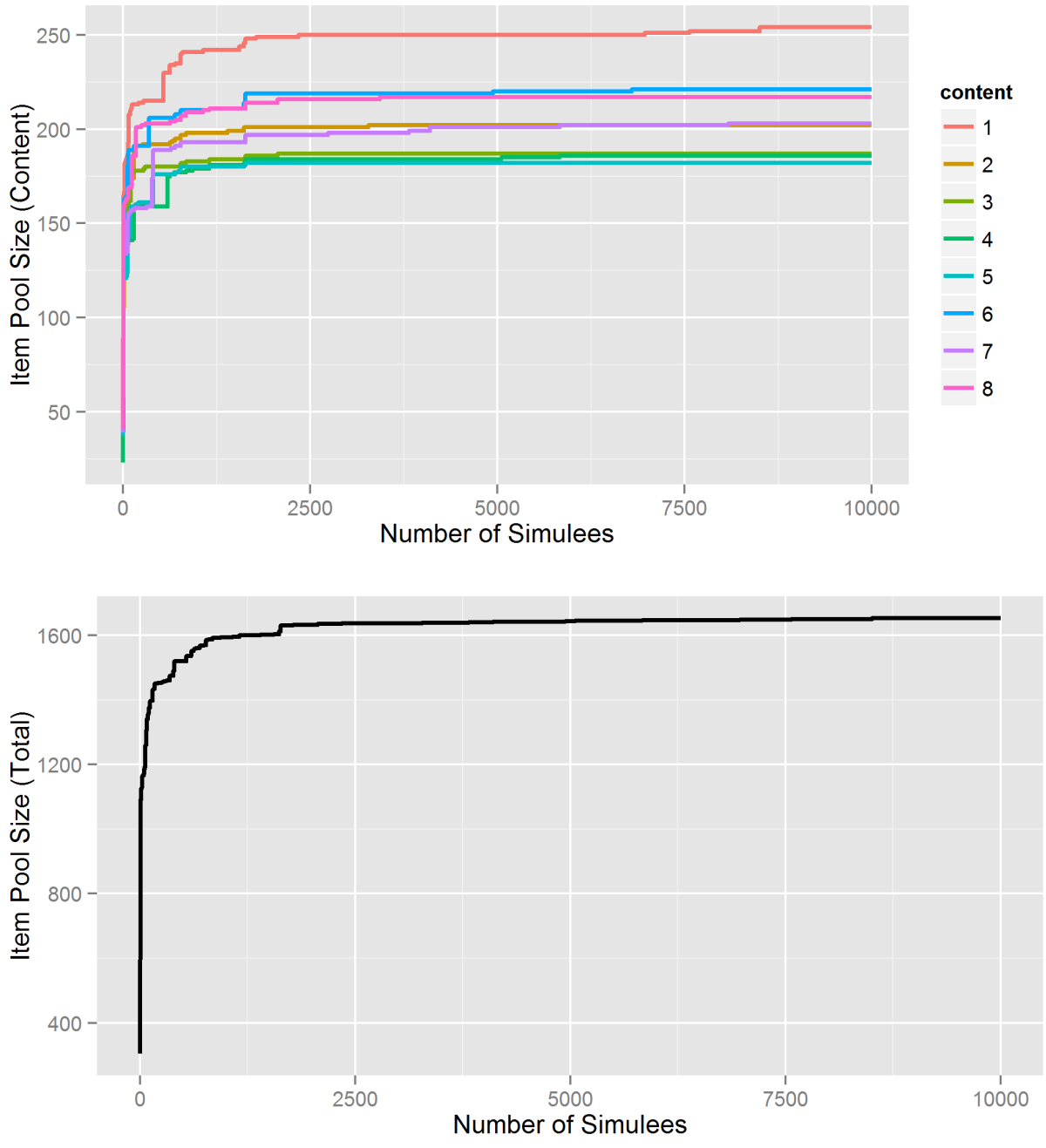

Figure F.1: Progress Plot for Ideal Item Pool with Fixed Bin Size 0.8

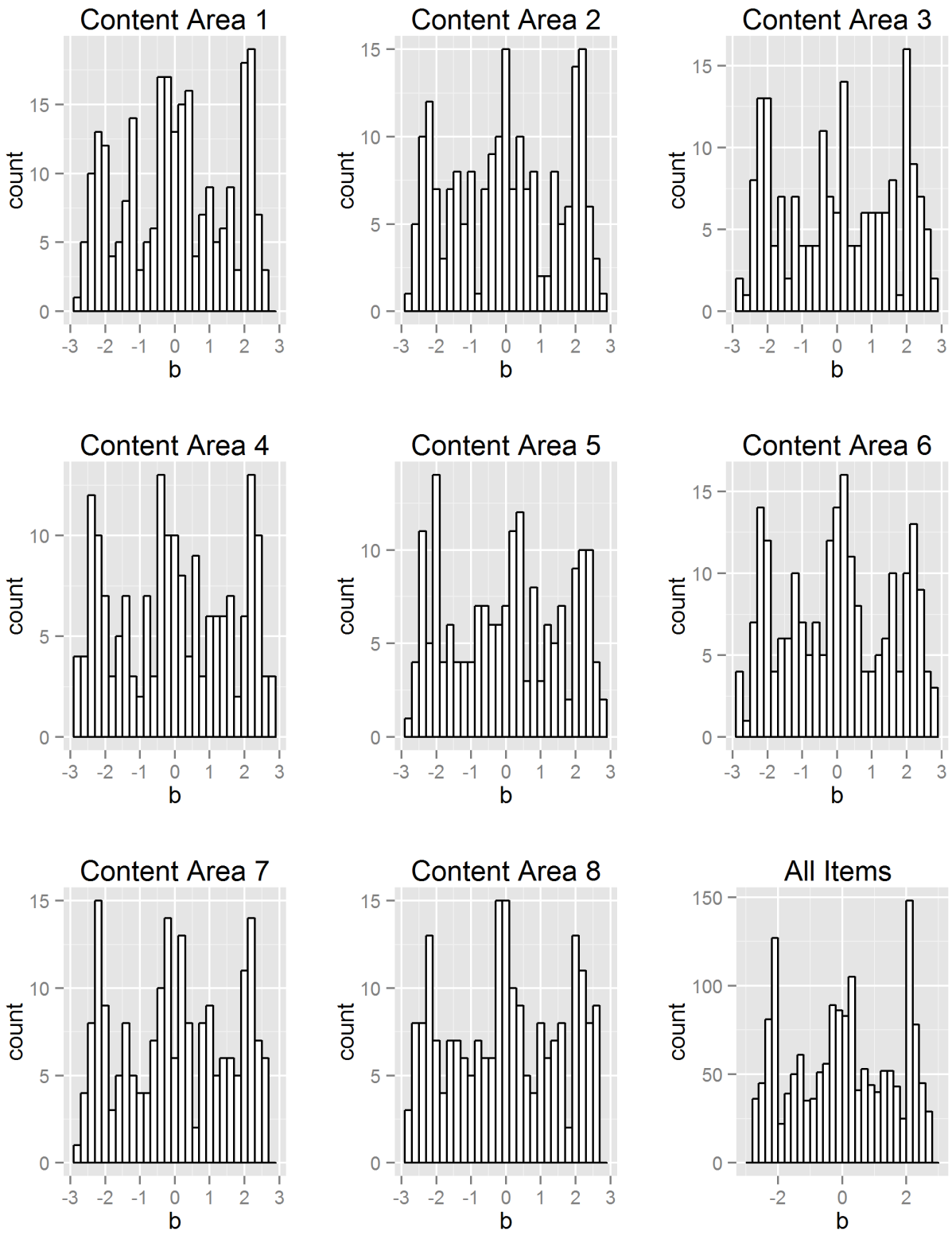

Figure F.2: Item Difficulty Distributions by Content Area for Ideal Item Pool with Fixed Bin Size 0.8

APPENDIX G

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 4

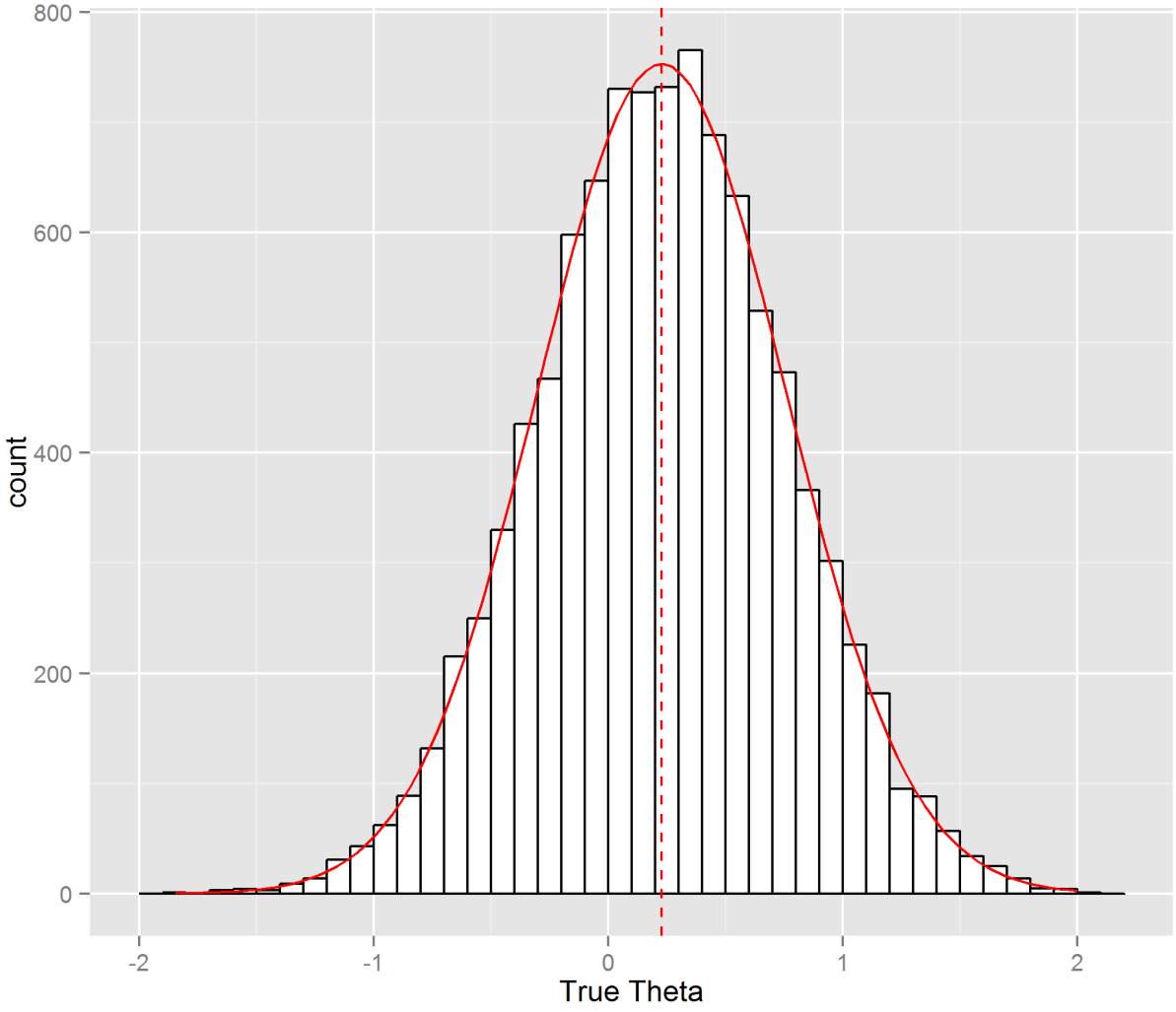

Figure G.1: True θ Distribution for Research Question 4

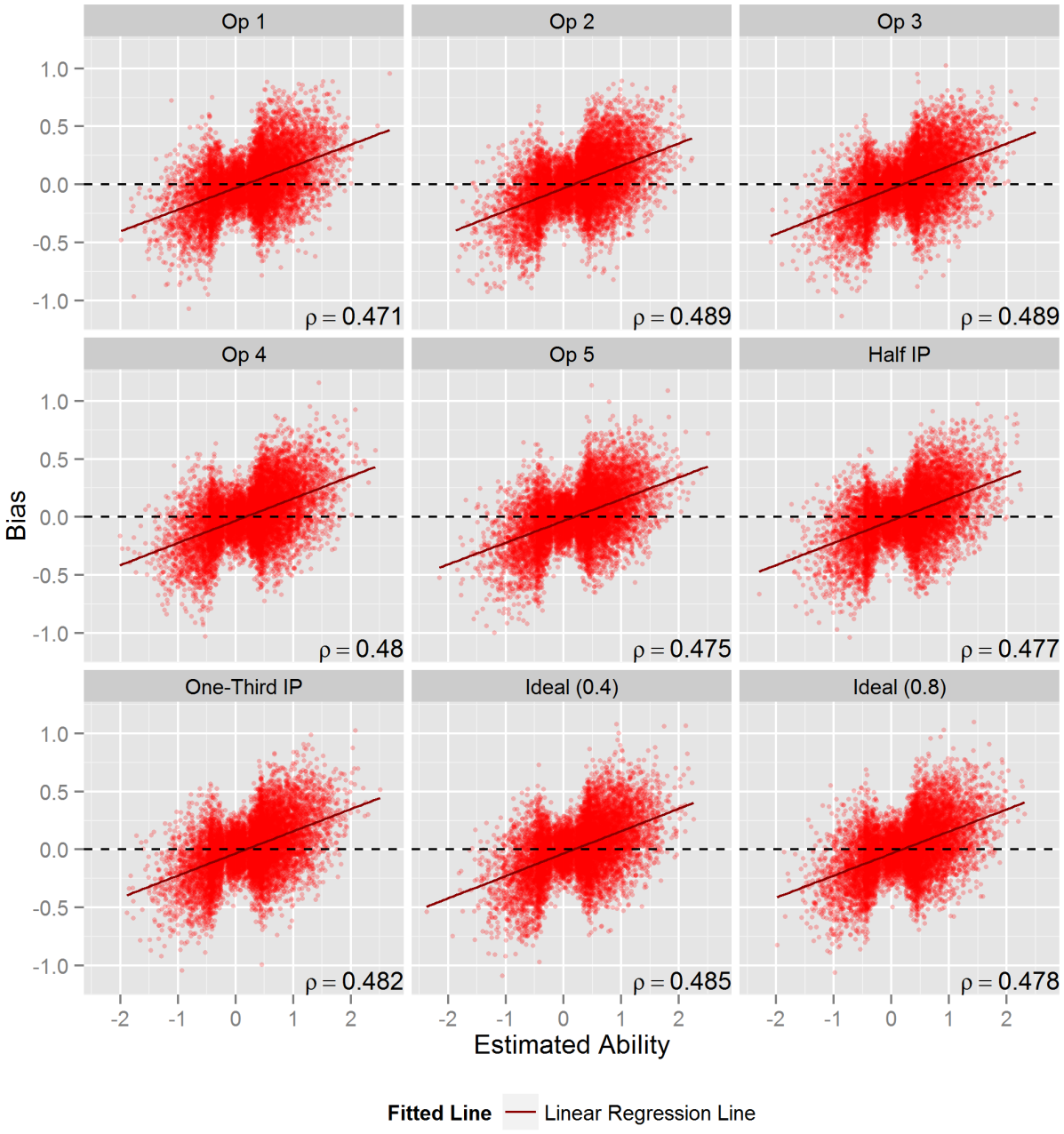

Figure G.2: The Relationship between Estimated Ability and Bias for each Item Pool Condition

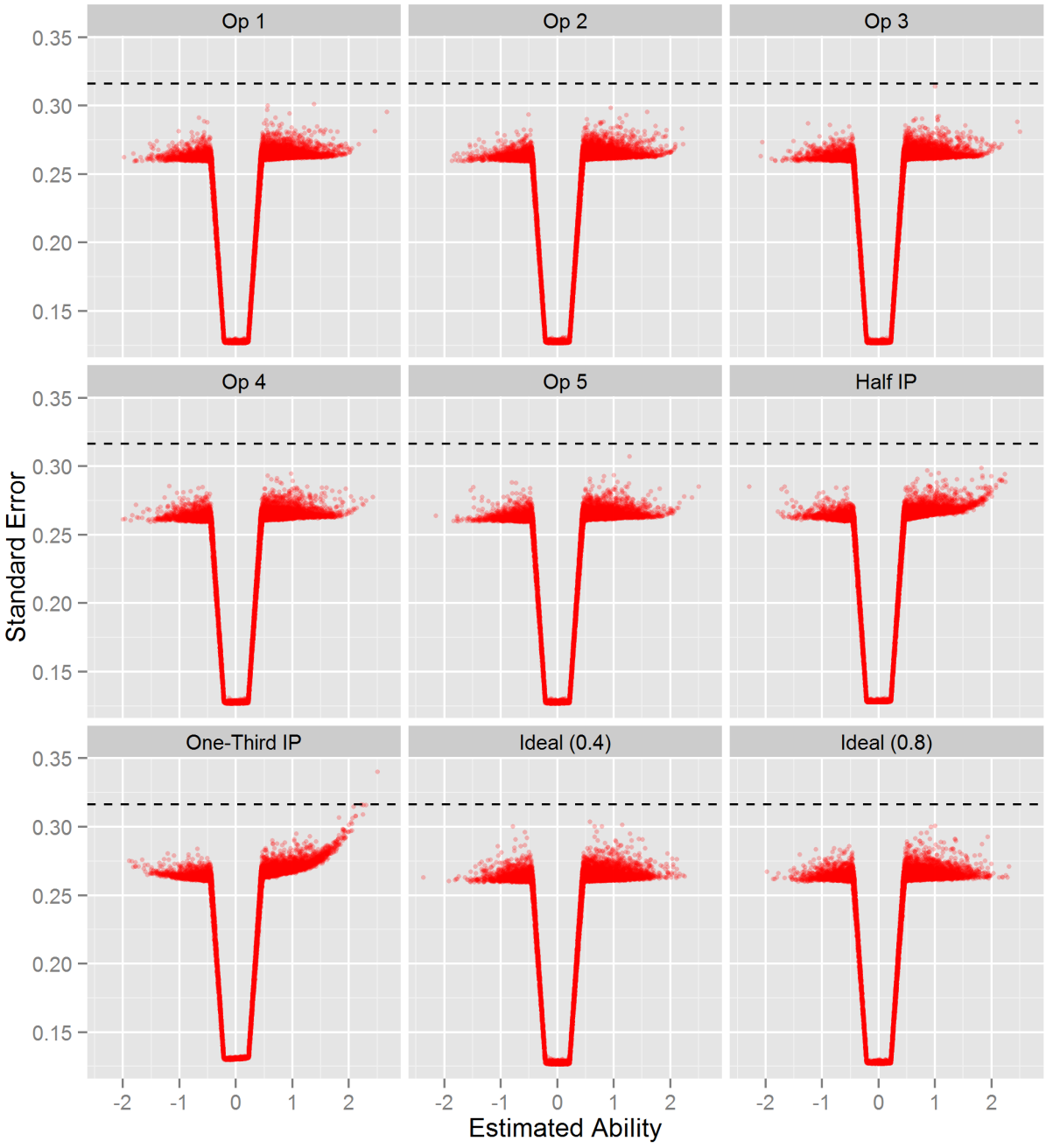

Figure G.3: The Relationship between Estimated Ability and Standard Error for each Item Pool Condition

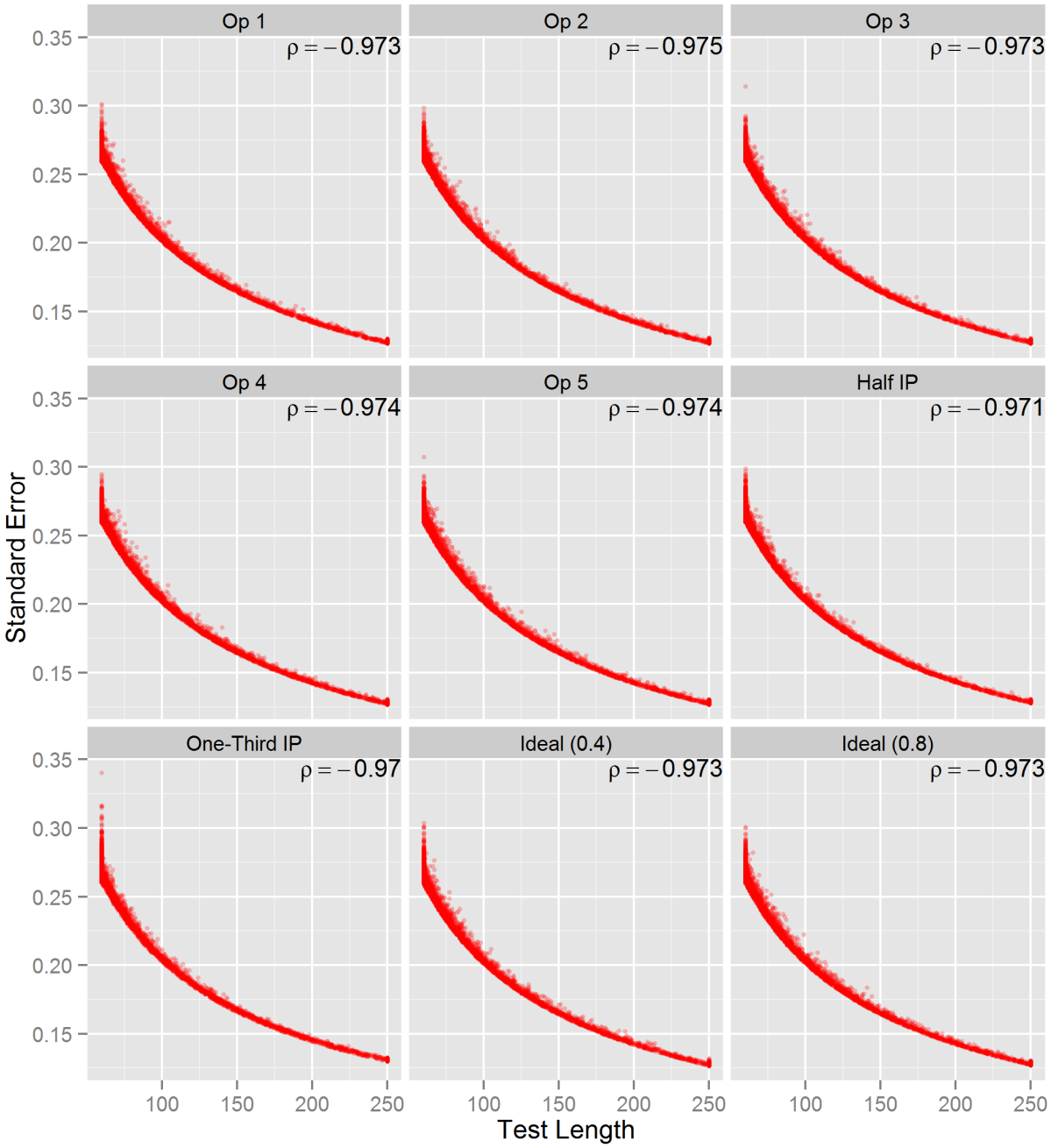

Figure G.4: The Relationship between Test Length and Standard Error for each Item Pool Condition

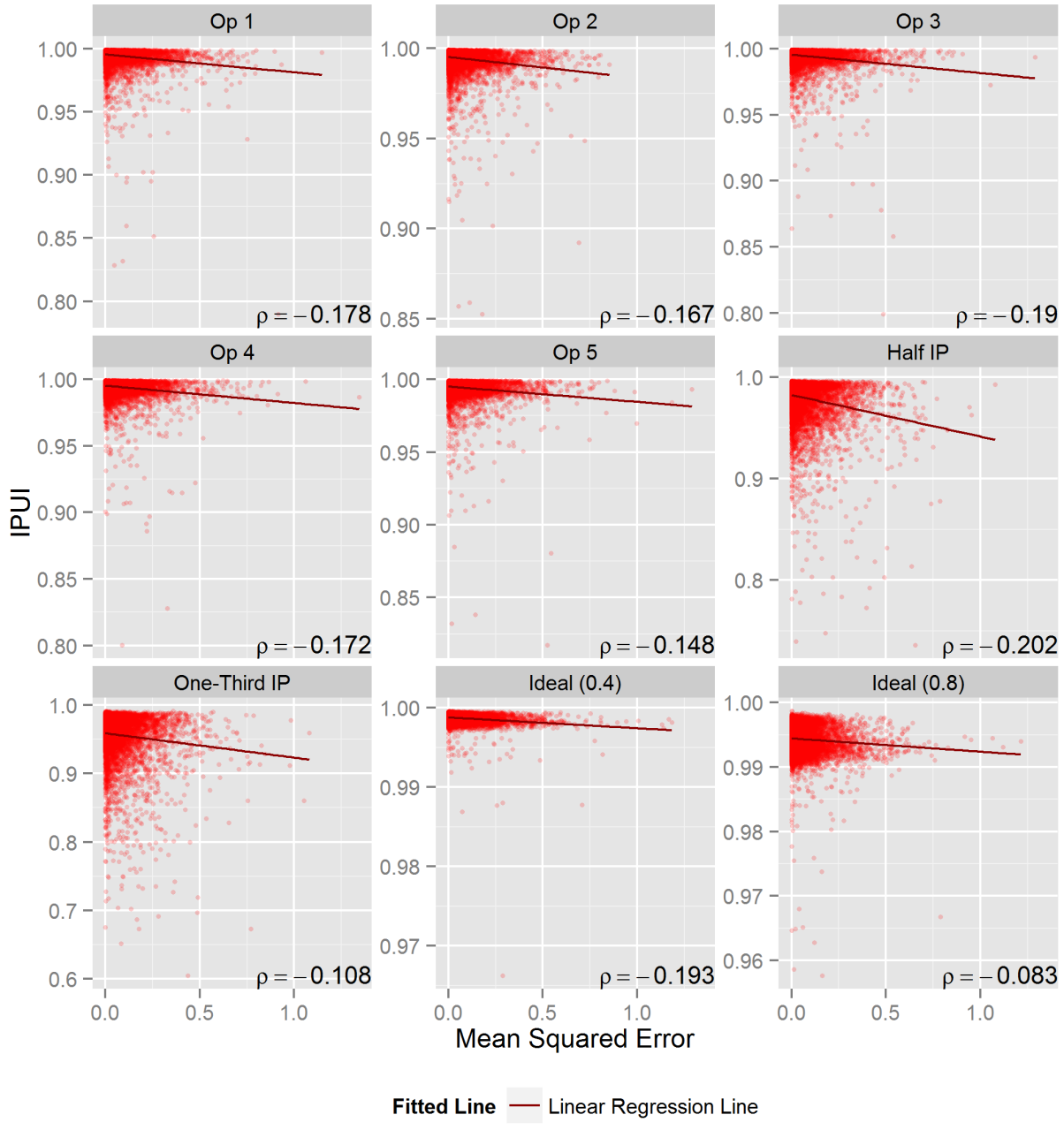

Figure G.5: The Relationship between IPUI and Mean Squared Error for each Item Pool Condition[1]

[1]Note that the x-axis scale for each sub-figure is different.

APPENDIX H

SUPPLEMENTARY FIGURES FOR RESEARCH QUESTION 5

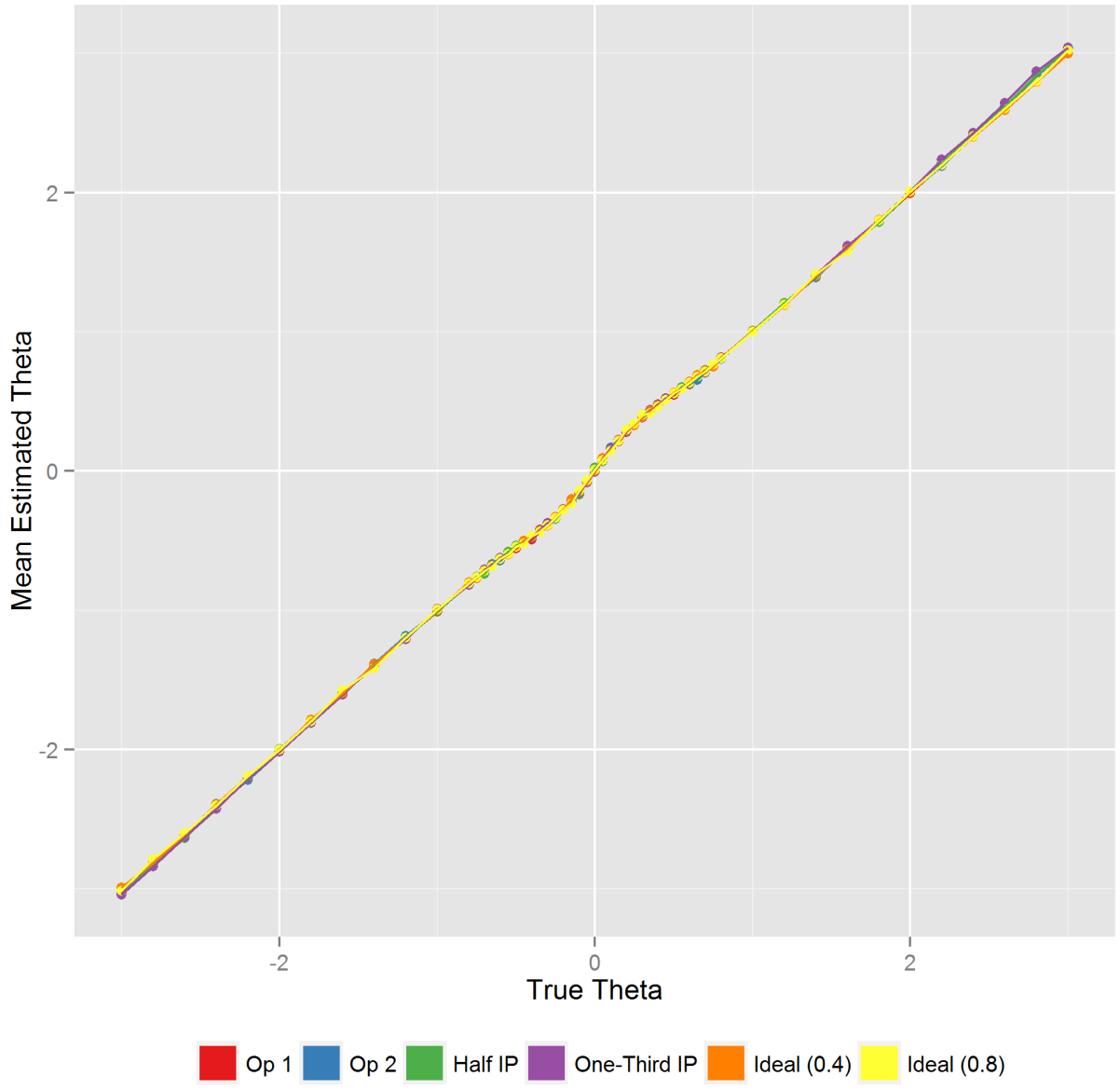

Figure H.1: The Relationship between True θ and Estimated θ for each Item Pool Condition

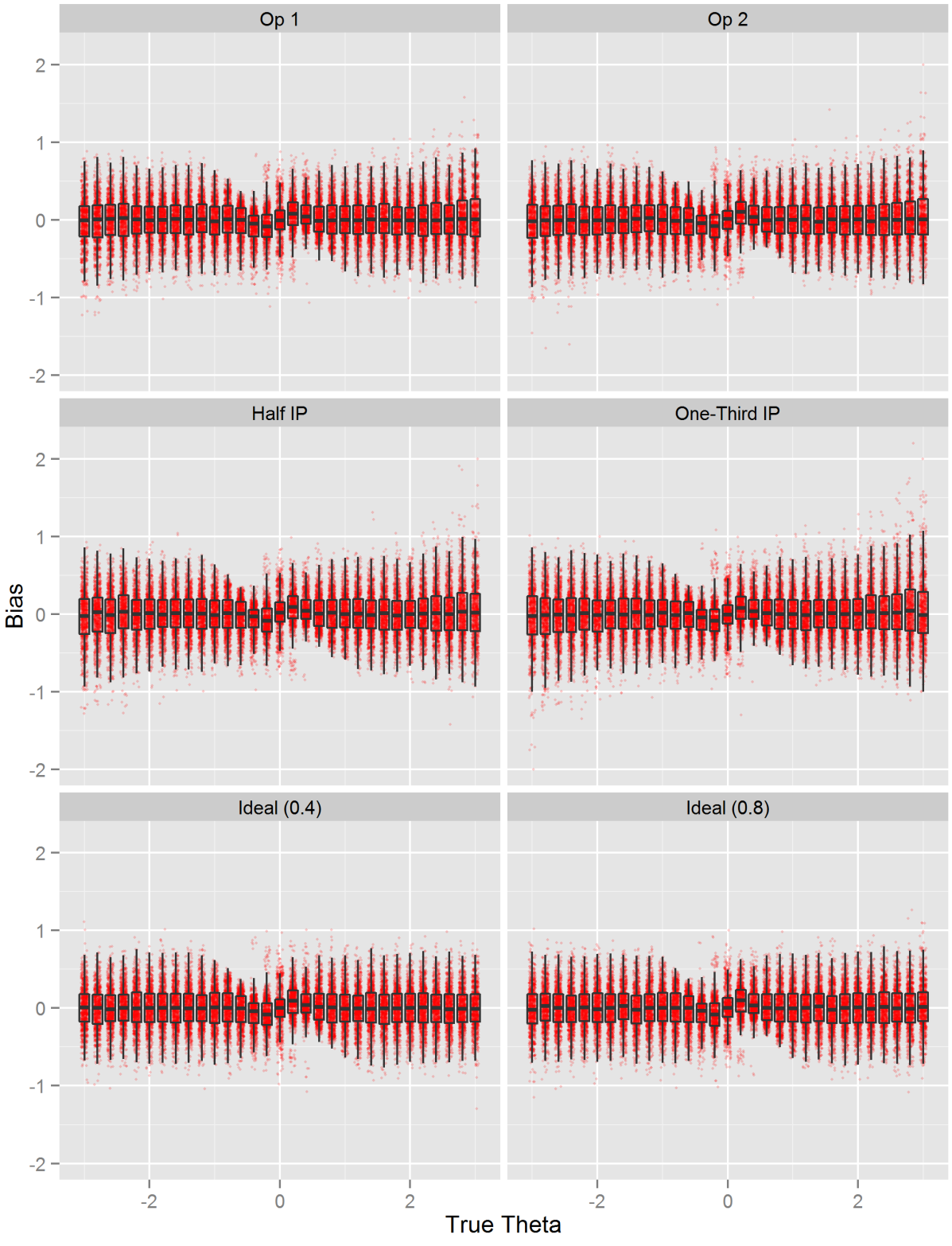

Figure H.2: The Bias Distribution at each True θ Value for each Item Pool Condition[1]

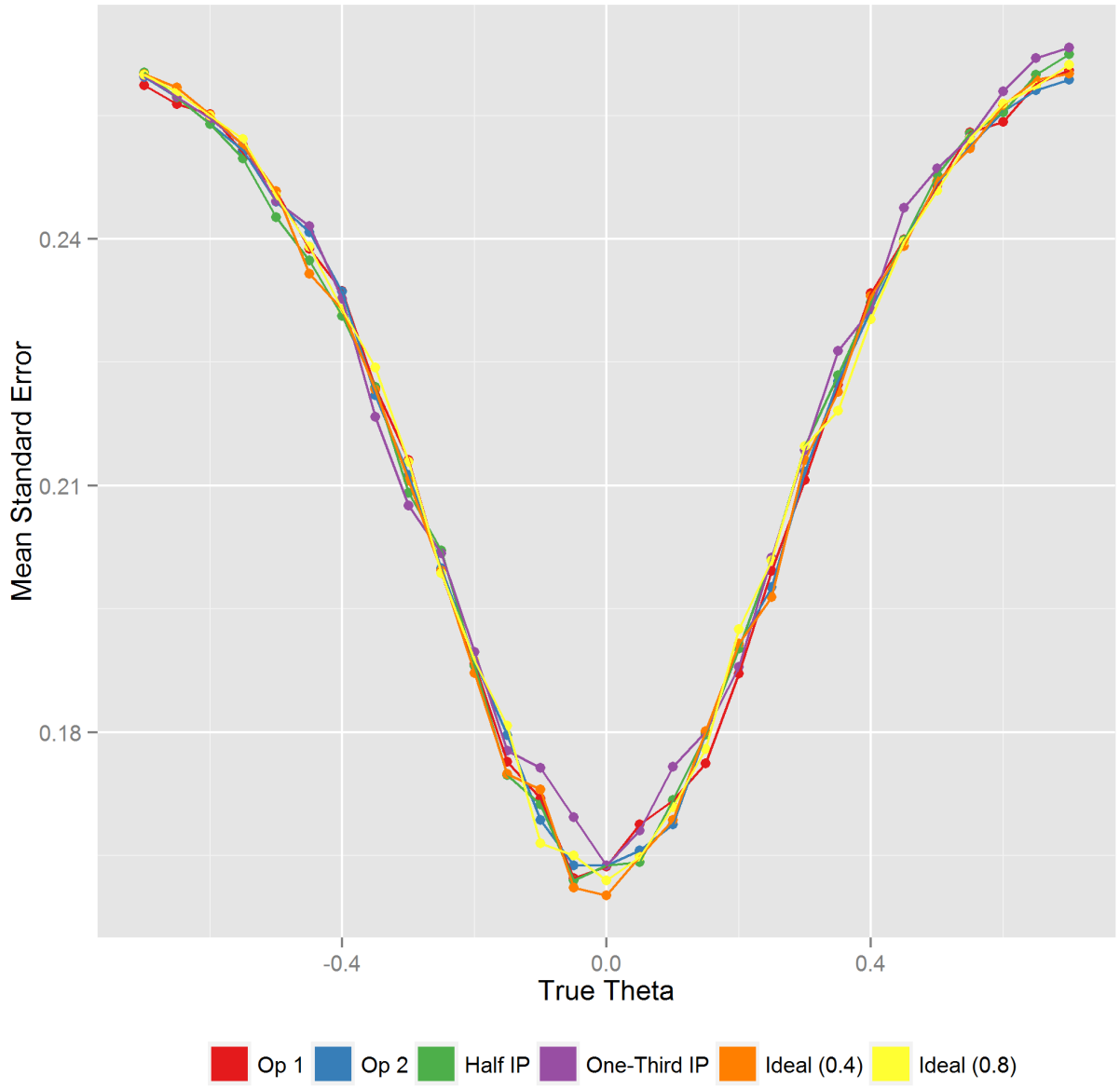

Figure H.3: Mean Standard Error Conditional on Restricted True θ Range for each Item Pool Condition

[1]For brevity, only some of the true θ values are displayed.

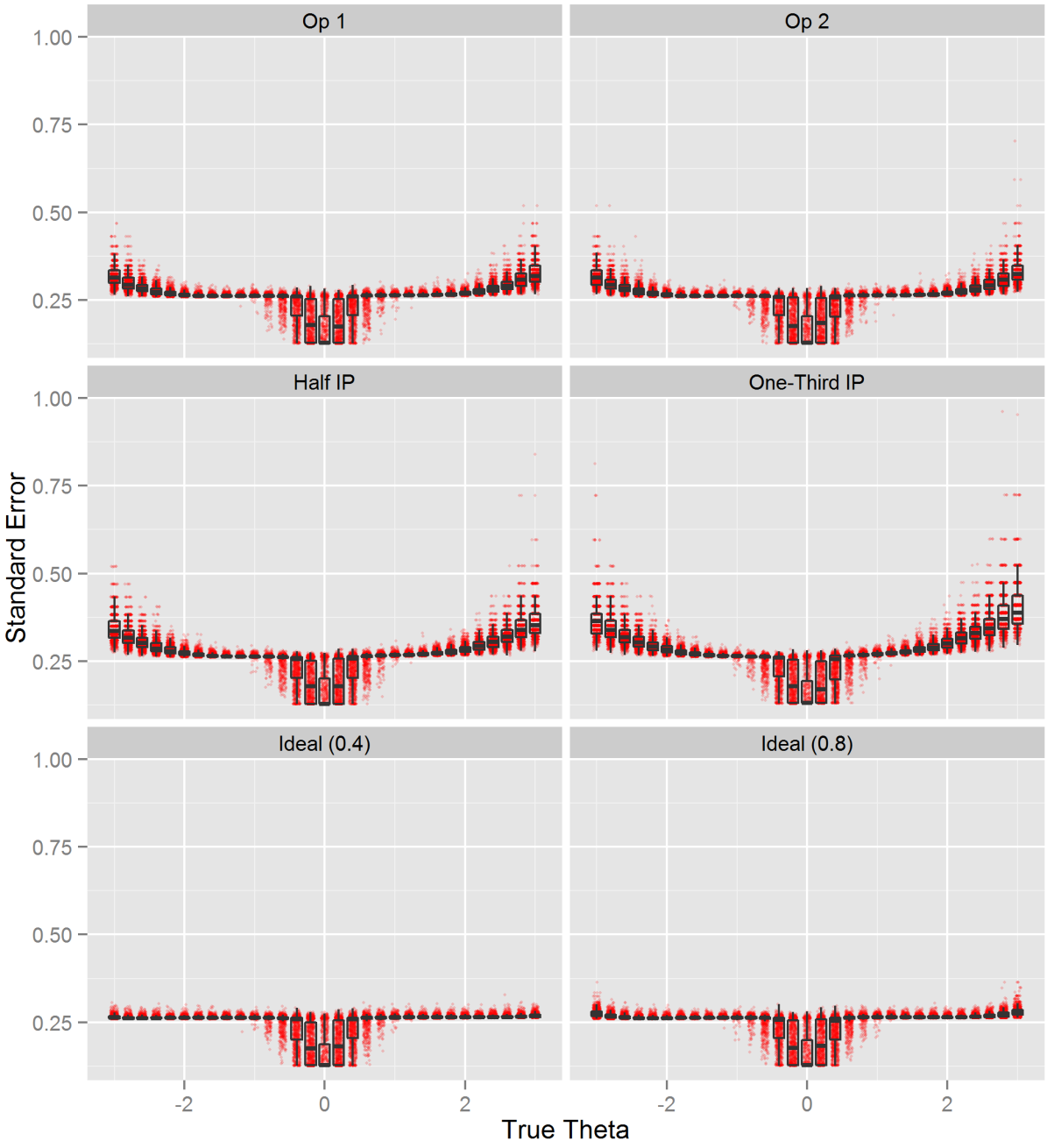

Figure H.4: The Standard Error Distribution at each True θ Value for each Item Pool Condition[2]

[2]For brevity, only some of the true θ values are displayed.

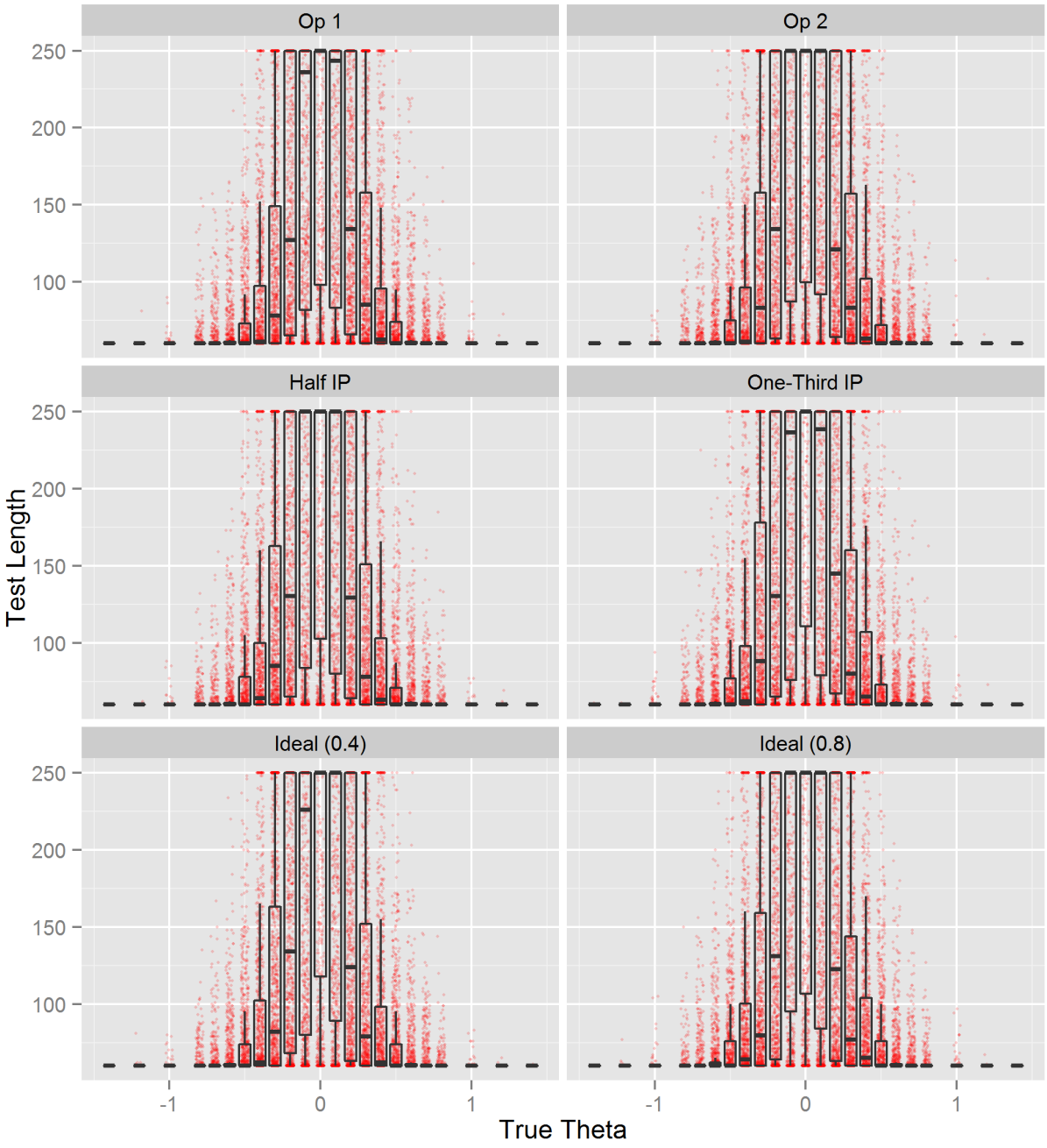

Figure H.5: The Test Length Distribution at each True θ Value for each Item Pool Condition[3]

[3]For brevity, only some of the true θ values are displayed. For θ values outside the [-1.5, 1.5] interval, the test lengths were all 60.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., ... Lai, J.-S. (2010). Development of a PROMIS item bank to measure pain interference. *PAIN*, *150*(1), 173–182. doi:http://dx.doi.org/10.1016/j.pain.2010.04.025

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, *34*(6), 438–452. doi:10.1177/0146621610370152

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303–310. doi:10.1177/014662168300700306

Belov, D. I. & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement*, *69*(4), 533–547. doi:10.1177/0013164409332224

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*(2), 137–149. doi:10.1207/s15324818ame0502_4

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. doi:10.1177/014662168200600405

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2010). Assembling an inventory of multistage adaptive testing systems. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 247–266). Springer.

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*(4), 333–341. doi:10.1177/01466210122032181

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. doi:10.1177/014662169602000303

Chang, H.-H. & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, *73*(3), 441–450. doi:10.1007/s11336-007-9047-7

Chang, H.-H. & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*(3), 1466–1488. doi:10.2307/30243674

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*(3), 241–255. Retrieved from http://apm.sagepub.com/content/24/3/241.abstract

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*(2), 129–145. doi:10.2307/1435342

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*(2), 369–383. doi:10.1348/000711008X304376

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi:10.1007/BF02310555

Davey, T. & Nering, M. L. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 165–191). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Eggen, T. J. H. M. & Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*(5), 713–734. doi:10.1177/00131640021970862

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*(5), 379–393. doi:10.1177/0146621606288890

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation*. Educational Testing Service.

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 edition, pp. 37–60). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, *5*(8).

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*(4), 361–8.

Gierl, M. J. & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36–50. doi:10.1111/emip.12018

Haley, S. M., Fragala-Pinkham, M. A., Dumas, H. M., Ni, P., Gorton, G. E., Watson, K., ... Tucker, C. A. (2009). Evaluation of an item bank for a computerized adaptive test of activity in children with cerebral palsy. *Physical Therapy*, *89*(6), 589–600.

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

Han, K. T. (2012). An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement*, *49*(3), 225–246. doi:10.1111/j.1745-3984.2012.00173.x

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*. doi:10.1177/0013164413517503

He, W. & Reckase, M. D. (2013). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*. doi:10.1177/0013164413509629

Hetter, R. D. & Sympson, J. B. (1997). Item-exposure in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: from inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Asociation.

Hildebrand, F. B. (1987). *Introduction to numerical analysis* (2nd edition). Mineola: NY: Courier Dover Publications.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.12000

Kim, J. K. & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*(4), 587–599. doi:10.1007/BF02294829

Kingsbury, G. G. & Wise, S. L. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicolgica: Revista de metodologa y psicologa experimental*, *21*(1), 135–156.

Kingsbury, G. G. & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: when is short too short? *International Journal of Testing*, *12*(4), 321–344. doi:10.1080/15305058.2011.643517

Leeuw, J. d. & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*(3), 183–196. doi:10.2307/1165071

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*(5), 857–874. doi:10.1177/0013164413486802

Lord, F. M. (1974). The relative efficiency of two tests as a function of ability level. *Psychometrika*, *39*(3), 351–358. doi:10.1007/BF02291708

Lord, F. M. (1975). Relative efficiency of number-right and formula scores. *British Journal of Mathematical and Statistical Psychology*, *28*(1), 46–50.

Lord, F. M. (1977a). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*(1), 95–100. doi:10.1177/014662167700100115

Lord, F. M. (1977b). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*(2). doi:10.2307/1434011

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*(2), 157–162. Retrieved from http://www.jstor.org/stable/1434513

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

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.

McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement*, *1*(1), 121–140. doi:10.1177/014662167700100119

Meijer, R. & Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194.

Millman, J. & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement*, *21*(4), 315–330. doi:10.2307/1434584

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287–304. doi:10.1207/s15324818ame0904_1

National Council of State Boards of Nursing. (2012). *NCLEX-RN examination detailed test plan for the National Council Licensure Examination for Registered Nurses item writer-item reviewer-nurse educator version*. National Council of State Boards of Nursing. Chicago, IL. Retrieved from https://www.ncsbn.org/2013_NCLEX_RN_Detailed_Test_Plan_Educator.pdf

Nering, M. L. & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). New York: McGraw-Hill.

Owen, R. (1969). *A Bayesian approach to tailored testing* (Report No. Research Bulletin No. 69-92). Educational Testing Service.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*(350), 351–356.

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.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. doi:10.1177/1073191111411667

R Core Team. (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

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.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, *52*(2), 127–141.

Reckase, M. D. & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361–373. doi:10.1177/014662169101500407

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327. Retrieved from http://www.jstor.org/stable/1435308

Rudner, L. M. (2010). Implementing the graduate management admission test computerized adaptive test. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 151–165). Springer.

Samajima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th edition, pp. 307–353). Westport, CT: ACE/Praeger Publishers.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. doi:10.1007/bf02294343

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: from inquiry to operation* (pp. 117–130). Washington, DC: American Psychological Asociation.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Report No. RR-94-05). Educational Testing Service. Princeton, New Jersey.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*(2), 151–166. doi:10.1177/014662169301700205

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1), 1–9.

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement*, *14*(2), 181–196. doi:10.2307/1434014

Vale, C. D. & Weiss, D. J. (1977). *A rapid item-search procedure for bayesian adaptive testing. research report 77-4* (Report No. Research Report 77-4).

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216.

van der Linden, W. J. (1998b). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*(3), 195–211. doi:10.1177/01466216980223001

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*(1), 81–99. doi:10.3102/10769986031001081

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer.

van der Linden, W. J., Veldkamp, B. P., & Reese, L. M. (2000). An integer programming approach to item bank design. *Applied Psychological Measurement*, *24*(2), 139–150. doi:10.1177/01466210022031570

Veldkamp, B. P. & van der Linden, W. J. (2010). Designing item pools for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (Chap. 12, pp. 231–245). New York: Springer.

Wainer, H. (2000). Introduction and history. 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, pp. 1–21). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135. doi:10.1111/j.1745-3984.1998.tb00530.x

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi:10.1007/bf02294627

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*(4), 17–27. doi:10.1111/j.1745-3992.1998.tb00632.x

Way, W. D., Steffen, M., & Anderson, G. S. (2002). Developing, maintaining, and renewing the item inventory to support 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. 143–164). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. doi:10.1177/014662168200600408

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

Weiss, D. J. & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*, *8*(3), 273–285. doi:10.1177/014662168400800303

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*(2), 21–30. doi:10.1111/j.1745-3992.2006.00054.x

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*(1), 5–21. doi:10.1177/0013164403258393

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement*, *38*(8), 614–631.