Effects of Item Bank Design and Item Selection Methods on Content Balance and Efficiency in Computerized Adaptive Reading Tests with Mix of Short and Long Passages Aligned to Common Core State Standards

Shudong Wang
NWEA

Liru Zhang
Delaware Department of Education

Paper presented at the annual meeting of the National Council on Measurement in Education. April 16-20, 2015, Chicago, IL

Send correspondence to:
Shudong Wang (Shudong.Wang@NWEA.org)
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206

Acknowledgement

We would like to thank Mike Nesterak from NWEA for his review and edit on this paper.

Effects of Item Bank Design and Item Selection Methods on Content Balance and Efficiency in Computerized Adaptive Reading Tests with Mix of Short and Long Passages Aligned to Common Core State Standards

Introduction

The Common Core State Standards (CCSS) are educational standards that include a set of consistent learning goals for what students should know and be able to do in English language arts/literacy and mathematics at each grade level across states. The ultimate goal of the CCSS is to ensure that all students have the skills and knowledge necessary to succeed in college, career, and life upon graduation from high school, regardless of where they live. One of the major purposes of the CCSS is to develop and implement a framework for common comprehensive assessment systems that measure student performance and provide teachers with specific feedback to help ensure students are on the path to success.

The Smarter Balanced Assessment Consortium (SBAC), as one of two state-led consortiums, is committed to ensuring that the assessment and instruction that embody the CCSS ensure that all students, regardless of disability, language, or subgroup status, have the opportunity to learn this valued content and to show what they know and can do. Because the CCSS contains a great deal of rationale and information about instruction and were not specifically developed for assessment, any assessment that claims to measure the CCSS has to provide evidences to prove it measures the CCSS. The major evidence includes the alignment between the CCSS and the assessment. The alignment between the SBAC assessment and the CCSS consists of two major components (Smarter Balanced Assessment Consortium, 2012), (1) the SBAC content specifications, and (2) the SBAC specifications for items/tasks. The latter is an extension of the former. The content specification is used to guide assessment development and the item/task specifications are intended to guide item and tasks developers in the future. Both specifications are intended to ensure that the assessment system accurately assesses the full range the standards. The other elements of the SBAC content structure are the test specifications (including a blueprint) that describe the make-up of the two assessment components, computer adaptive test (CAT) and performance assessment, and how their results will be combined and reported. The SBAC capitalized on the precision and efficiency of the computerized adaptive

testing (CAT) for both the mandatory summative assessment and the optional interim assessments. In general, test efficiency is defined as the mean number of items required to reach a certain reliability level. For a fixed length CAT, it means the accuracy of examinees' ability estimation from the test.

In the process of test development and evaluation, the validity is the most important consideration. There are five sources of validity evidence specified in the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). These evidences are: (a) test content, (b) response process, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. The test validation process in K-12 assessments relies heavily on content validation procedures (Kane, 2006; Lissitz & Samuelsen, 2007). Because test content alignment is at the core of content validity and an aligned test must include items and tasks requiring the highest level of cognitive complexity prescribed by the test claims and targets, the SBAC has adopted a Cognitive Rigor Matrix (CRM, Hess, 2004) for its assessment program (Smarter Balanced Assessment Consortium, 2012). The CRM draws from two widely accepted measures to describe cognitive rigor: Bloom's (revised) Taxonomy of Educational Objectives (Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, Wittrock, 2001; Bloom, Engelhart, Furst, Hill, Krathwohl, 1956) and Webb's Depth-of-Knowledge Levels (Webb, 1997, 1999). The CRM has been developed to integrate these two models as a strategy for analyzing instruction, for influencing teacher lesson planning, and for designing assessment items and tasks.

Reading Comprehension (RC) is one of the most measured subjects in K-12 education. The RC as part of English Language Arts and Literacy specified by the CCSS is one of the mandatory accountability measures. At the heart of the CCSS for English Language Arts (ELA) and Literacy is the shift of instruction to center on text. The standards focus on the growing complexity of texts (or passages) and using evidence from texts to present analyses and well-defended claims. Instead of asking questions that can be answered solely from prior knowledge and experience, students are expected to answer a range of text-dependent questions that require inferences based on the text (Coleman and Pimentel, 2012). In RC test, a short passage usually associate with one item (discrete item) and a long passage usually associate with a group of items

(testlet). The term "testlets" refers to a group of items related to a single content area and has been defined differently by different people. It can be a packet of test items that are administered together (Wainer & Kiely, 1987), a coherent groups of multiple choice items based on a common stimulus (Bradow, Wainer, Wang, 1999), a bundle of items that share a common stimulus (Wang & Wilson, 2005c), a subset of items in a test that have a common structural element (Brandt, 2008), and an aggregation of items on a single theme (Ip, 2010). The "modules" (Sheehan & Lewis, 1992) refers to preconstructed sets of items that are packaged and administered as intact units. In practice, there is nothing precluding the use of these labels to describe a set of items that includes everything ranging from a cluster of items to a set of performance exercises (Luecht & Nungester, 1998; Luecht, 2002a; Luecht et al., 2006; Luecht & Sireci, 2011). The stimulus materials are used in the SBAC ELA assessments include traditional passages, audio presentations, and scenarios for students to react to (Smarter Balanced Assessment Consortium, 2012).

Unless there are infinite numbers of items in the bank that have adequate coverage of test properties required in a test specification, regardless of which method is used, the danger of exhausting the item bank before certain test attributes can be satisfied always exists and could lead to less optimal measurement precision. Because the major advantage of CAT compared to a linear test is its efficiency in terms of the accuracy of examinee's ability estimation or the short test length, using CAT to provide optimal measurement of examinee's RC ability is very appealing. However, in practice, in order to provide valid and reliable measurement with a minimal number of items using CAT, additional factors, such as test security (item exposure rates) and content balance are also important concerns for assessment practitioners (Davey & Parshall, 1995). The CAT design can vary from adapting at individual item level as classical CAT to adapting at group of items level as computer-adaptive multi-stage testing (ca-MST, Luecht & Nungester, 1998; Luecht & Sireci, 2011). In general, the CAT design that adapts at individual item level and without any items selection constraints due to requirements of test validity and administration (such as test content representation and test security), is the most efficient CAT design (baseline CAT) in terms of accuracy of examinee's ability estimation. However, in practice, a valid CAT test is not baseline CAT because item selection constraints always exist for test validity and administration concerns. It is crucial to understand there is an inevitable trade-

off between CAT efficiency and practice constraints (or validity). A higher efficiency test without careful considerations of constraints is usually associated with low test validity for the given CAT pool.

The two major factors that could reduce CAT efficiency discussed here are adaptive design and content constraint. First, any modification of baseline CAT design, such as design that mixes both CAT and ca-MST, will lead to less efficient test design compared to a baseline CAT. Figures 1 to 6 show a variety of designs that incorporates adaptation method at both item and item group levels. The CAT design that adapts at individual item level (either with or without constraints) is shown in Figure 1. The Figure 2 represents the ca-MST. Since items within testlet or module in ca-MST are linearly administrated, an alternative design named ca-MST(i-CAT) is shown in Figure 3. The only difference between ca-MST and ca-MST(i-CAT) is the way of items are administrated within a testlet (internal items). All internal items of a testlet are adaptive rather than linearly administrated and this internal adaptation is abbreviated as i-CAT. The other type of design is the mix-design that mixes ca-MST/ca-MST(i-CAT) and CAT designs. Figures 4 and 5 depict these two designs in which CAT test items are administrated first and ca-MST/ca-MST(i-CAT) items are administrated second. The natural extension of a mix-design is a design that reverses the administration order of CAT and ca-MST/ca-MST(i-CAT) or a multiple mix of CAT and ca-MST/ca-MST(i-CAT) as delineate in Figure 6. As mentioned above, in general all designs that employ adaptation at the item group level than at the individual item level will lead to less efficient test in a statistical sense. Although all internal items of ca-MST(i-CAT) design are adapting at the individual, they are also adapting at the testlet level. The efficiency of this design is still less than that of CAT. In theory, the rank order of efficiency of these designs form high to low could be (1) CAT, (2) CAT and ca-MST(i-CAT),(3) CAT and ca-MST, (4) ca-MST(i-CAT), and (5) ca-MST.

The second major factor that could hinder CAT efficiency is item selection constraints, such as constraints on item content and statistical properties. Any item selection method that deviates from the maximizing item information method will reduce the efficiency in selection of items that match a test taker's provisional ability estimation. Because item selection in CAT is sequential by nature, different approaches have been proposed to deal with its inability to use backtracking (van der Linden, 2005). The major approaches include the proportion content

balance method (Kingsbury & Zara, 1991), the weighted-deviations methods (WDM) (Stocking and Swanson, 1993), and the shadow-test approach (STA) (van der Linden, 2000; 2005, chap. 9; van der Linden & Reese, 1998). Assuming the CAT pool has adequate coverage over content specification and aligns to the CCSS, DOK concerns, the nature of RC items within a testlet design still requires more consideration on how to efficiently arrange RC items in the test. Passage length, passage complexity, the number of items, range of difficulty of items, and the mean or median of items within a passage could also have impact on the efficiency of CAT. Because there are so many concerns in real CAT applications for RC test alignment to CCSS, this paper couldn't focus on all the concerns mentioned above and treats them as fixed. The major purpose of this paper is to investigate the effects of item selection methods that includes testlet and individual item selections and different pool distributions on the content validity and efficiency of mixed CAT and ca-MST(i-CAT) designs.

Methods

1. Design of Study

A Monte Carlo (MC) simulation method is used to evaluate the relationship between efficiency of CAT in terms of accuracy of ability estimation and content validity in terms of content coverage in this study. In this CAT simulation study, one balanced factorial experimental design and both descriptive methods and inferential procedures are used. The manipulated three independent variables include: (1) distribution of item bank (C: 1 to 4), (2) individual item selection algorithm (I: 1 to 3), and testlet selection algorithm (T: 1 and 2). The three dependent variables that describe average or overall accuracy of person estimation in this study are (1) biases, (2) standard errors (SEs), and (3) root mean square errors (RMSEs). These three average dependent variables indexes are

$$Bias(\hat{\theta}) = \frac{1}{R}\frac{1}{T}\frac{1}{N}\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{n=1}^{N}(\hat{\theta}_{rtn} - \theta_{rt}) \quad (1)$$

$$SE(\hat{\theta}) = \frac{1}{R}\frac{1}{T}\sum_{r=1}^{R}\sum_{t=1}^{T}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{rtn} - \frac{1}{N}\sum_{n=1}^{N}\hat{\theta}_{rtn})^2} \quad (2)$$

$$RMSE(\hat{\theta}) = \frac{1}{R}\frac{1}{T}\sum_{r=1}^{R}\sum_{t=1}^{T}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{rtn} - \theta_{rt})^2} \quad . \quad (3)$$

Where θ_{rt} is the true ability of simulees in rth replication, which was used to generate responses in the simulation and R is number of replication (total 100 replications). There are the 4 equally spaced conditional true ability levels or thetas that range from –2 to 0 in increments of 0.5 and 4 equally spaced true ability levels that range from 0 to 2 in increments of 0.5. The T is the number of true ability (total 8 true abilities). A CAT was simulated for 1000 simulees at each of the 8 ability parameter points for each of replications and the $\hat{\theta}_{rtn}$ is the estimated ability for nth number of simulees. Besides, the average bias, SE, and RMSE, the conditional bias, SE, and RMSE are also calculated to describe the distribution of accuracy of ability estimation method, they are:

$$Bias(\hat{\theta}|\theta_T) = \frac{1}{R}\frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N}(\hat{\theta}_{rn} - \theta_r) \quad (4)$$

$$SE(\hat{\theta}|\theta_T) = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{rn} - \frac{1}{N}\sum_{n=1}^{N}\hat{\theta}_{rn})^2} \quad (5)$$

$$RMSE(\hat{\theta}|\theta_T) = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_{rn} - \theta_r)^2} \quad . \quad (6)$$

Because the MC study is really a statistical sampling experiment with an underlying model, the number of replications in this MC study is the analogue of sample size. In this study, in order to have adequate power for the statistical tests in the Monte Carlo study to detect effects of interest, the conditional indices had been replicated 1000 times at each given ability level and the simulation results from each three-way ANOVA design for each of overall indexes had been replicated 100 times. Because the N is equivalent to sample size for conditional theta, so N x

R=1000 x 100 = 100,000 are the analogue of total sample size for MC study. The experimental design is used in the analyses of the overall indices, the significance of a statistic is tested, and the empirical sampling distributions for the statistics are generated. The 4C x 3 I x 2T completely crossed analysis of variance (ANOVA) designs were used.

Besides the inferential statistic for 3 dependent variables, the descriptive statistics results, such as that item exposure rate and differences of item attributes (sub-content and DOK) between target test specification and real test results are also provided.

2. CAT Design

Because the primary goal of this study is to assess the effects of item selection methods and different pool distributions on the content validity and efficiency of mixed CAT and ca-MST(i-CAT) design, many variables related to the purpose of this study but not focused in this study are treated as fixed, such as item bank size, the passage Lexile measure, and the word counts in passage. The item response model used in this study is the Rasch model (Rasch, 1960). The reason to use the Rasch model rather than Rasch testlet model (Jiao, Wang & He, 2013) in this study is that the Rasch testlet model has not been used in real applications. The CAT employs a fixed test length design and the test length is 40 items. In this mixed CAT and ca-MST(i-CAT) design as shown in Figure 5, the first 20 items are discrete items and last 20 items are testlet items. There are total 4 testlets in the tests and there are 5 items in each testlet. Table 1 presents the plan of the design.

2.1 Item Bank

For the purpose of this study, the simulated item bank consists of two type items, discrete items and testlet items. The discrete items represent short passages and the testlet items imitate long passages. Because the real size of the item bank for a given grade and subject used by the SBAC is unknown at this time, this study targets the item bank size that tends to be larger than real one for given grade and subject. In this bank, there are 900 discrete items and the generated b parameters of discrete item difficulty are from normal distribution b ~N(0, .5). There are also 200 testlets and the numbers of items within testlets have the range from a minimum of 5 items to a maximum of 10 items. The total number of items in the banks is 2380 and the only

difference between item banks is the distribution of content and DOK. The Figure 7 to 8 displays the distributions of the numbers of items within testlets and all item difficulties in the bank. For testlet items, the mean of the distributions of testlets (mt) is from mt ~N(0, 1) and the mean of the distributions of testlets is illustrated in Figure 9. It is assumed that the test specification requires the est to have 4 sub-content areas and 3 level of DOK. Based on this assumption, across discrete and testlet items in the bank, items have been generated to meet the specifications and Table 2 shows the characteristics of the item bank and target test properties. From Table 2, it is clear that the proportion of item sub-content in the bank matches the proportion of the test property. However, the proportion of item DOK in the bank does not match the proportion of the test property. Besides sub-content area and DOK concerns, the item difficulties are also generated to have slight correlations with item DOKs.

Although proportions of sub-content and DOK in item banks are known, the proportions of sub-content and DOK within each of the testlet type items are not exactly known even though the distributions of sub-content and DOK follows those in Table 2. The reason of this is due to the fact that the number of items within testlets has a range from 5 to 10. For example, a testlet that has 5 items guarantees to have one item from each of four sub-content areas, but the 5^{th} item may come from any one of four sub-content areas. The same situation applies to DOK. Unless the number of items within a testlet is very large, both sub-content and DOK distributions of items within testlets will unlikely match the distributions of the test specifications exactly.

2.2 Method for Selecting the First Item

In this study, the first item for any given examinee has difficulty value that is 0.5 logit lower than examinee's true ability.

2.3 Ability Estimation Method

The two ability estimation methods used are modified Owen Bayesian estimation (OBE, Divgi, 1986) and Maximum Likelihood estimation (MLE). The OBE provides provisional ability estimation to select items and the MLE provide final scores.

2.4 Item Selection Method

Item selection methods used in this study are sequential. This means the selection is in a hierarchical order. The item selection method includes two components; the first is testlet selection (T) and second is individual item selection (I).

There are total 3 individual item methods. The first is an information based method (IM), the second is an information plus sub-content balance (ICM) method, and the last is the second method plus DOK constraint (ICDM).

There are total 2 testlet selection methods that include (1) using criteria to select testlet and (2) not using criteria to select testlet. For using the testlet criteria method, all three individual item selection methods set the testlet selection as the highest order among all criteria. The testlet selection criterion used is the absolute difference between the median (b_{med}) of item difficulties within the testlet and the provisional ability estimation ($\hat{\theta}$) to select next testlet and item, i.e., if $\left|\hat{\theta} - b_{med}\right| < 0.3\,(\text{level } 1) \text{ or } 0.5\,(\text{level } 2)$ in logit, then the next item will be selected from this testlet. Once the testlet has been selected, the rest of next 4 items also has to come from the same testlet. Not using the testlet criteria method, the $\left|\hat{\theta} - b_{med}\right|$ is set to equal to 5 logits so that the range of difference is so large and it can be ignored.

Regardless of selection methods, all items selected have to satisfy the test design. This means that all of the first 20 items have to be discrete items even if other items from testlets may be optimal in terms of information, content, and DOK. This similar logic is true for testlet items. The last 20 items have to come from testlets even if some discrete items may have optimal values in terms of information, content, and DOK. Each of these methods follows the same common procedure that provides three tiers of structure:

Tier 1. Selecting a group items (such as 5 to 15 items) from the item bank based on criteria level one. If none item can be found, then go to tier 2.

Tier 2. Selecting a group items from the item bank based on criteria level two. If none item can be found, then go to tier 3.

Tier 3. Selecting a group items from the item bank that have the best (by sorting) values of criteria in tier 2.

All three tiers use randomization strategy (McBride & Martin, 1983; Kingsbury & Zara, 1989) to control for item exposure by randomly selecting an item from a given group of items

mentioned above. Because the number of items within testlets is larger than the number of items for a given testlet in a test, testlet items are adaptively selected for a given testlet using information, content, and DOK accordingly.

For the IM method, the level one or two criterion is item Fish information (FI). Because the maximum FI for Rasch model is 0.25, this study uses 0.235 as the level one and 0.216 as the level two information criteria. Any item that has FI >=0.235 will be selected in tier 1 and any item that has FI >=0.216 will be selected in tier 2. If both tier 1 and tier 2 can't find any item, then a group of items that have the largest FI values (less than 0.216) will be selected for a given provisional ability.

For the ICM method, besides FI criterion used in the IM method, the additional proportion content balance method (Kingsbury & Zara, 1991) is used in both tiers 1 and 2 for both level one and level two. In tier 3, if there is no item required for the sub-content that has the largest the difference between expected and real proportions in bank at any given stage of test, then the ICM will drop this constraint requirement. This guarantees that the test will not stopped but will lead to a content deviation of the test based on the test specification.

Compared to ICM method, the ICDM method adds an additional DOK balance methodology that insures the proportions of DOK levels in the test meet DOK requirements of content specifications. The DOK balance method is used in both tiers 1 and 2 for both level one and level two. In tier 3, if there is no item required for the DOK level that has the largest difference between expected and real proportions in the bank at any given stage of test, then the ICDM will drop this constraint requirement. This guarantees that the test will not stopped but lead to a DOK deviation of test of the test specifications. It is worth mentioning that for both ICM and ICDM methods, the test deviations from content or DOK can be diminished as the size of a well-balanced item bank increases.

All simulations and statistical analyses are conducted by using SAS (SAS Institute Inc. 2013).

Results

Two major focuses of this study are test efficiency in terms of the accuracy of examinee ability estimation for given item bank and test validity that is measured by how well the test matches the test specification.

1. Accuracy of Test

The CAT test efficiency is embodied within the accuracy of examinee's ability estimation or the short test length compared to a linear test. In this study, the accuracies of examinees' ability estimation are expressed in terms of bias, SE, and RMSE under different simulation (or experiential) conditions (4 item banks and 3 selection methods).

1.1 Overall Accuracy

Overall indices of bias, SE, and RMSE are used as dependent variables in this study. The three independent variables are the distribution of the item bank (B), individual item selection method (I), and testlet selection method (T). The SAS procedure Proc GLM has been used for all ANOVA.

Table 3 exhibits the results of the three-way ANOVA of bias. Using $\alpha = 0.01$ for each hypothesis tested, two main effects: I (Item) and T (testlet), and one two-way interaction effect: I x T were statistically significant. The proportion of total variation accounted for by the effect being tested are measured by Semipartial Eta-Square ($\eta 2$) (Cohen, 1988; Maxwell, 2000). The results show that both the main effects (I and T) and the two-way interaction effect (IT) accounted for large variance in bias. Following the advice of Cohen (Cohen, 1988), the effect size in terms of $\eta 2$ had been classified as: (a) no effect ($\eta 2 < 0.0099 \approx 0.01$), (b) small effect ($0.01 < \eta 2 < 0.0588 \approx 0.06$), (c) medium effect ($0.06 < \eta 2 < 0.1379 \approx 0.14$), and (d) large effect ($\eta 2 > 0.14$). Among these significant effects (I, T, and IT), the I effect accounted for most of the variance (95.99%) with the T being next. 1.21 % of the variance in the bias for θ estimate is due to the effect of the T. The interaction effect of IT is demonstrated in figure 10. Except for I=1(IM), both the ICM and ICDM methods, using testlet criteria (T=1) will lower the bias more

than not using it (T=2). Figure 11 displays mean biases (simple effect) under different simulation conditions. Again, in general, using testlet criteria (T=1) will lower the bias than not using it (T=2).

Table 4 illustrates the results of the three-way ANOVA of SE. Using $\alpha = 0.01$ for each hypothesis tested, two main effects, I (Item) and T (testlet), and one two-way interaction effect, I x T, were statistically significant. Among these significant effects (I, T, and IT), the I effect accounted for most of the variance (95.99%) with the T effect being next. 1.87 % of the variance in the bias for θ estimate is due to the effect of the T. Figure 12 displays the interaction effect of IT. It shows, across different individual item selection methods, using testlet criteria (T=1) will lower the bias more than not using it (T=2).

Table 5 illustrates the results of the three-way ANOVA of SE. Using $\alpha = 0.01$ for each hypothesis tested, two main effects, I (Item) and T (testlet), and one two-way interaction effect, I x T, were statistically significant. Among these significant effects (I, T, and IT), the I effect accounted for most of the variance (95.66%) with the T being next. 2.12 % of the variance in the bias for θ estimate is due to the effect of the T effect. Figure 13 explains the interaction effect of IT. It shows, across different individual item selection methods, using testlet criteria (T=1) will lower the bias more than not using it (T=2).

In general, both I and T factors have an impact on all dependent variables in this study. The factor I has the most influence on bias, SE, and RMSE and accounts for more than 95% of the total variance. The factor T has statistically significant but less effect on bias, SE, and RMSE than factor I. For ICM and ICDM, using a testlet criteria will reduce bias, SE, and RMSE.

1.2 Conditional Accuracy

Three conditional indices that were computed at each of eight true theta levels are bias, SE, and RMSE. Figure 14 depicts the conditional accuracies of ability estimation in terms of bias, SE, and RMSE under four different item banks (B), testlet (T) and individual item (I) selection methods across eight examinees' true ability.

First, ability estimation of different item banks and item selection methods have almost no bias at the middle range of the ability scale and are slightly biased at the extremes of the ability scale. Since the final ability estimation method is MLE, the direction of all these biases is

called “outward” direction, which means that bias is under-estimated at low ability level and over-estimated at high ability level. The bias figures also reveal that the IM is the best method comparing to the ICM and ICDM across different items banks. Among all testlet selection methods, not using a testlet criterion tends to lower the bias for both ICM and ICDM methods.

Second, all ability estimations have smaller SE at the middle range of ability and larger SE at the extremes of ability range. There is almost no difference of SE among different distribution of item banks. Among all item selection methods, IM has the lowest value of SE and there is almost no difference of SE between ICM and ICDM across different banks. Among all testlet selection method, not using a testlet criterion tends to inflate the SE for both ICM and ICDM methods.

Lastly, all ability estimations have smaller RMSE at the middle range of ability and larger RMSE at the extremes of the ability range. There is almost no difference of RMSE among different distribution of item banks. Among all item selection methods, IM has the lowest value of RMSE and there is almost no difference of RMSE between ICM and ICDM across different banks. Among all testlet selection method, not using testlet criterion tends to inflate the RMSE for both ICM and ICDM methods.

In general, the results for conditional indexes are consistent with the results of overall indices. The largest differences among dependent variables come from item selection methods and the IM is the most accurate method. Using testlet criterion tends to increase the accuracy for both ICM and ICDM methods.

2. Validity of Test

The validity measured in this study is how well the test meets test specification presented in Table 2 under test property. The CRM specified in this study include sub-content and DOK.

2.1 Content Coverage

The figure 15 depicts the distributions of difference of sub-contents coverage between mean and targeted percentages across different item banks and item selection methods. For example, the “-5 Difference of % Content” for sub-content area one (say, C1) means that the test has 20% of items in sub-content one and is 5% less than the targeted percentage which is 25%.

Regardless of the item bank and item selection method used, at the middle range of ability scale, the match between the test and test specifications is almost perfect because the differences of all four sub-content coverages between mean and targeted percentages are almost zero. As the ability range moves away from center to the two extremes, the mismatches between test and test specification increase. The mismatch between of sub-content areas between the test and the test specifications is due to the fact that both sub-content and DOK distributions of items within testlets are not matched with the distributions of test specifications exactly. Because the IM item selection method does not put constraints on either content and DOK, among three item selection methods, the IM has the worst match at the two extremes for item bank 1 and 2. The testlet selection method has a small impact on sub-content and not using the testlet criterion will reduce the difference of sub-content coverage between mean and targeted percentage.

Overall, the matches of sub-content areas between test and test specifications are very good. This is mainly due to the fact that the test specifications match item bank specifications in terms of percentages of sub-content areas.

2.2 DOK Coverage

Figure 16 delineates the distributions of difference of DOK coverage between mean and targeted percentages across different item banks and item selection methods. It is clearly shown that matches of DOK coverages between test and test specifications across different item banks and item selection methods is much worse than that of sub-content areas.

This is due to two main reasons. First, the DOK matches between item banks and test specifications are not as good as matches of sub-content areas as shown in Table 2. For example, for the first two item banks, the percentages of DOK Level 1 are 0.2, while the percentage of DOK Level 1 in test specification is 0.4. Additionally, the percentages of DOK level 3 are 0.3, while the percentage of DOK level 1 in the test specification is 0.2. For both level 1 and level 3, the percentages of DOK levels in the test are either higher or lower than that in the item banks. The similar situations exist for DOK level 2, except for item bank 3 and 4. In this case, both percentages of DOK level 2 in the item bank 3 and 4 are equal to the percentages of DOK level 2 in test specifications. The second reason is that among all three item selection methods, satisfying DOK requirement in the test is always the lowest priority.

Because of the first reason, regardless of item selection method used, for item bank 1 and 2, all DOK level 1 are under-represented and all DOK levels 2 and 3 are over-represented in the tests. By the same token, all DOK level 3 are under-represented in tests for item banks 3 and 4; all DOK level 1 are over-represented for item banks 3 and 4; and DOK level 2 has close matches for item banks 3 and 4.

The impact of testlet selection method on DOK is not as clear as the impact of it on sub-content. This may be due to the fact that DOK has lower priority in terms of selection compared to the sub-content area.

The results here indicate that the quality of the item bank has direct impact on the quality of the test and the deficiency in the test could reflect the limitations of the item bank.

2.3 Item Exposures

The third validity issue of CAT is item exposure. The item exposure rate (IER) for any given item in an item bank is the proportion of the number of CAT administrations to the number of examinees. Figures 17 to 22 display the distributions of average (over 100 replications) conditional (at each of eight true thetas from -2 to 2 logits) IER in the item bank one (B=1) across item selection methods (M=1, 2, 3) for different testlet selection (T=1, 2). The horizontal axis represents the item ID number (from 1 to maximum 2380) in the bank and the vertical axis indicates the IER, and the # in IE# denotes the true theta, i.e., IE1= -2.0, IE2= - 1.5, IE3= - 1.0, IE4= -0.5, IE5=0.5, IE6=1.0, IE7=1.5, IE8=2.0. The first 900 items are discrete items and the rest of items are testlet items.

Under T=1(using testlet criterion), for any given item selection method, across true thetas, the discrete items have much smaller IERs than some of testlet items even though both types of items used the same randomization method to control for item exposure. The IERs become smaller as true thetas are closer to the middle of range (theta=0) and the IERs become larger as the true thetas are farther from the middle. The item selection methods ICM and ICDM have better control over IERs than method one IM method. For item banks two to four, the same situations exist.

Under T=2 (not using testlet criterion), IERs are very small regardless of distribution of item bank and individual item selection methods. It is clear that not using a testlet level selection constraint during item selection will dramatically improve the item exposure rate.

In general, the randomization method is very effective on controlling item exposure for discrete items and the IM method results in the worst item exposure when using testlet level selection.

Discussion and Conclusions

Content validity and efficiency of CAT is closely related to test design, adequacy of the item bank, and item selection algorithm used in a reading test that consists of both short and long passages. Until recently, researchers have paid little attention to the CAT achievement test that deals with mixed discrete and type items. In this study, the effects of item selection methods with item banks that align to CCSS on validity and efficiency of CAT are investigated.

Results across different test or item selection methods show that although the distribution of the item bank has little effect on bias, there is a practically significant effect on SE and RMSE. And the item selection method has also practically significant effect on SE and RMSE. The method (IM) that does not put constraint on sub-content and DOK has lowest SE and RMSE. And the methods (ICM and ICDM) that put item selection constraints with either sub-content or DOK, or both have approximately a quarter percent increase in SE and RMSE across the different item banks. This means that the content validity gain is due to efficiency loss of CAT.

Additionally, the distribution of item sub-content and DOK in the banks has a direct impact on the quality of test validity across item selection methods. The result of the study demonstrates that when an item bank closely aligns to test specification, the quality of test content validity improves. The relationship between the distribution of item difficulty in the banks and accuracy of ability estimation can be seen in figure 23. It is clear that the number of items with moderate item difficulty has the highest accuracy of ability estimation and as item difficulty becomes either more difficult or easier at the two extremes, the accuracy decreases.

Compared to the efficiency (or accuracy) of test, the distribution of content characteristics of items in the bank and the item selection method have important impact on test validity in terms of coverage of sub-content and DOK. The more constraints in the item selection will result more valid test results. The only testlet level selection constraint applied in this study is the distance between the median of item difficulties within testlet and provisional ability estimation. This criterion excluded many testlet items that might be selected to satisfy the need of sub-content and DOK. Unless there is very large and balanced item bank supporting the strong demand for test content validity requirements, for a given limited item bank, the more testlet level constraints applied, the less chance the test will satisfy the content requirements. In this study, there are 900 discrete items and 200 testlets that content at least five items. Even with this number of items, the tests are still not perfect aligned to the tests' specifications. Besides the concern of a content match, testlet items also have higher item exposure rates. Imagine having more testlet level attributes such as passage types applied to CAT tests specifications. Including both information and literature passages requirements could easily have doubled the size of the bank used in this study. Given the practical limitation in test development, it would be unlikely for any test organization to have money and resources to build such a large item bank. So the suggestion here is to limit the testlet level constraints to as few as possible.

Similar to the individual item selection method, the testlet selection method has the second largest the impact on accuracy of ability estimation. Removing testlet selection constraints in item selection will decrease accuracy for both ICM and ICDM methods across item bank, but will increase accuracy for IM methods. Removing testlet selection constraints also improves the validity of the test in terms of sub-content, but the improvement is diminished in terms of DOK.

In general, including testlet type item in CAT result in significant increase in item bank demands and complicates the item selection process. One way to reduce the burden on testlet item demands in a reading test is to mix short and long passage.

References

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

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., Wittrock, M.C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Pearson, Allyn & Bacon.

Bloom, B.S. (Ed.). Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay Co Inc.

Bradow, E. T., Wainer, H., & Wang, X (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51-70). Princeton, NJ: IEA-ETS Research Institute.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ: Erlbaum.

Coleman, D. and Pimentel, S. (2012). *Revised publishers' criteria for the Common Core State Standards in English Language Arts and literacy, Grades 3-12*. Student Achievement Partners. Supported by National Governors Association, Council of Chief State School officers, Achieve, Council of the Great City Schools, and National Association of State Boards of Education.

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

Divgi, D. R. (1986). *On Inaccuracies In Owen's Approximation for the Bayesian Ability Estimate*. CNA Research Memorandum 86-46. Retrieved from https://www.cna.org/sites/default/files/research/2786004600.pdf

Hess, K. (2004). *Applying Webb's Depth-of-Knowledge (DOK) Levels in reading*. [online] available: http://www.nciea.org/publications/DOKreading_KH08.pdf.

Ip, E. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34,* 467-482.

Jiao, H., Wang, S., Kamata, A. (2007). Modeling local item dependence with the hierarchical

generalized linear model. In E. V. Smith & R. M. Smith (ed.), Rasch Measurement: Advanced and Specialized Applications. JAM press.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*, 186-203.

Kane, M. (2006). Content-related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

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

Kingsbury, G. G., & Zara, A. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive testing. *Applied Measurement in Education, 4*, 241–261.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.

Luecht, R. M. (1998a, April). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R.M. (1998b). Computer assisted test assembly using optimization heuristics, *Applied Psychological Measurement, 22*, 224–236.

Luecht, R. M. (2002a, February). *An automated test assembly heuristic for multistage adaptive tests with complex computer-based performance tasks*. Invited paper presented at the Annual Meeting of the Association of Test Publishers, Carlsbad, CA.

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

Luecht, R., & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Research Report). College Board. Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2011-12-review-models-for-computer-based-testing.pdf

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for the uniform CPA examination. *Applied Measurement in Education, 19*, 189–202.

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

Maxwell, S. E. (2000), "Sample Size and Multiple Regression Analysis," *Psychological Methods, 5*, 434–458.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

SAS Institute Inc. (2013). SAS® 9.4 Guide to Software. Updates. Cary, NC: SAS Institute Inc.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16*, 65–76.

Smarter Balanced Assessment Consortium. (2012). *SMARTER Balanced Assessment Consortium General Item Specifications*. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf (accessed February 19, 2015).

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement. 42*, 283–302.

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

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

Wang, W.-C., & Wilson, M. (2005c). The rasch testlet model. *Applied Psychological Measurement, 29(2)*, 126–149.

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

Webb, N. (1997). Research Monograph Number 6: *Criteria for alignment of expectations and assessments on mathematics and science education*. Washington, D.C.: CCSSO.

Webb, N. (August 1999). Research Monograph No. 18: *Alignment of science and mathematics standards and assessments in four states.* Washington, D.C.: CCSSO.

Table 1. Plan of Design

Dependent Variables		Independent Variables	
Conditional	Overall	Distribution of Item Bank (B)	Item Selection Method (M)*
Bias	Bias	1, 2, 3, 4	1 (IM), 2 (ICM), 3 (ICDM)
SE	SE	1, 2, 3, 4	1 (IM), 2 (ICM), 3 (ICDM)
RMSE	RMSE	1, 2, 3, 4	1 (IM), 2 (ICM), 3 (ICDM)

*: See section 2.4 for detail information

Table 2. Characteristics of Item Bank and Targeted Test Property

Item Bank	% of DOK Level			% of Content Area				$r_{b\text{-}DOK}$
	1	2	3	1	2	3	4	
1	0.2	0.5	0.3	0.25	0.25	0.25	0.25	0.3
2	0.2	0.5	0.3	0.25	0.25	0.25	0.25	0.6
3	0.5	0.4	0.1	0.25	0.25	0.25	0.25	0.3
4	0.5	0.4	0.1	0.25	0.25	0.25	0.25	0.6
Test Property	0.4	0.4	0.2	0.25	0.25	0.25	0.25	

Table 3. Results of Three-Way ANOVA of Average Bias

Source	df	*F*	*p*	Semipartial Eta-Square η^2
Main Effects				
B (Bank)	3	1.32	0.2652	0.0000
I (Item)	2	44.50	<.0001	0.9599
T (testlet)	1	29.06	<.0001	0.0121
Interaction Effects				
B x I	6	1.23	0.2886	0.0000
B x T	3	0.17	0.9186	0.0000
I x T	2	18.39	<.0001	0.0063
B x I x T	6	0.40	0.8789	0.0000
Error	2376			

Table 4. Results of Three-Way ANOVA of Average SE

Source	df	*F*	*p*	Semipartial Eta-Square η^2
Main Effects				
B (Bank)	3	0.22	0.8798	0.0000
I (Item)	2	75646.6	<.0001	0.9599
T (testlet)	1	2944.92	<.0001	0.0187
Interaction Effects				
B x I	6	0.94	0.4629	0.0000
B x T	3	0.65	0.5855	0.0000
I x T	2	497.28	<.0001	0.0063
B x I x T	6	0.32	0.9294	0.0000
Error	2376			

Table 5. Results of Three-Way ANOVA of Average RMSE

Source	df	*F*	*p*	Semipartial Eta-Square η^2
Main Effects				
B (Bank)	3	0.20	0.8995	0.0000
I (Item)	2	77123.9	<.0001	0.9566
T (testlet)	1	3422.68	<.0001	0.0212
Interaction Effects				
B x I	6	0.95	0.4546	0.0000
B x T	3	0.59	0.6218	0.0000
I x T	2	595.11	<.0001	0.0074
B x I x T	6	0.32	0.9284	0.0000
Error	2376			

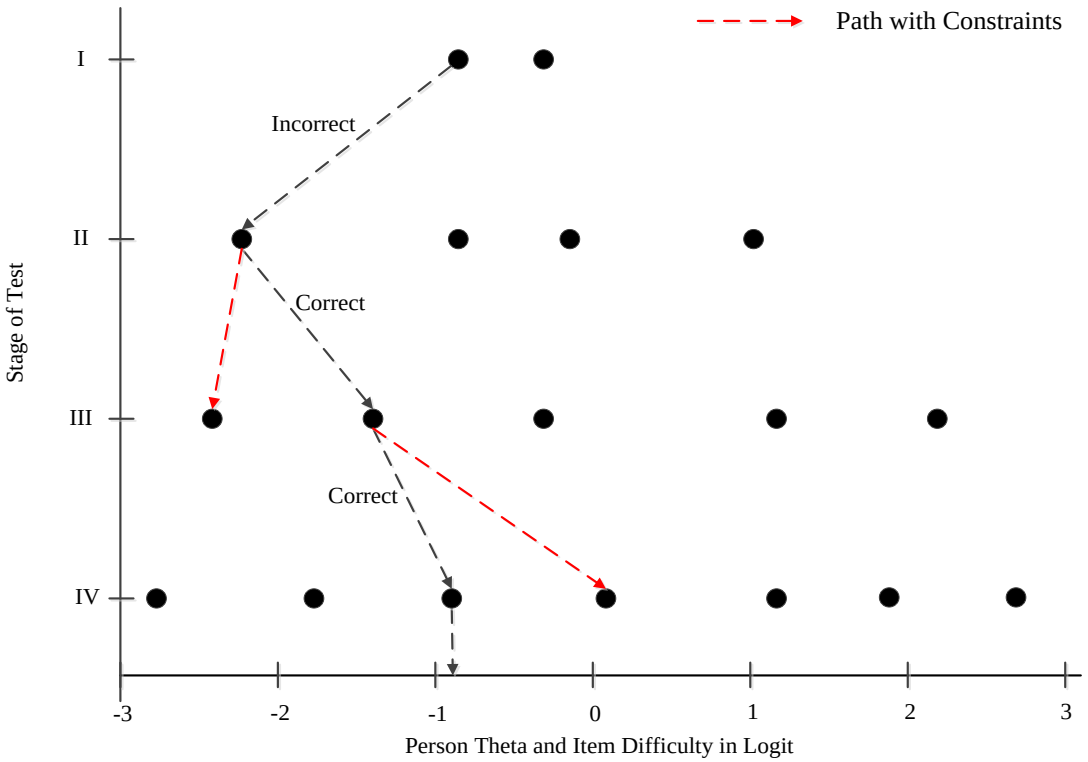

Figure 1. CAT Design

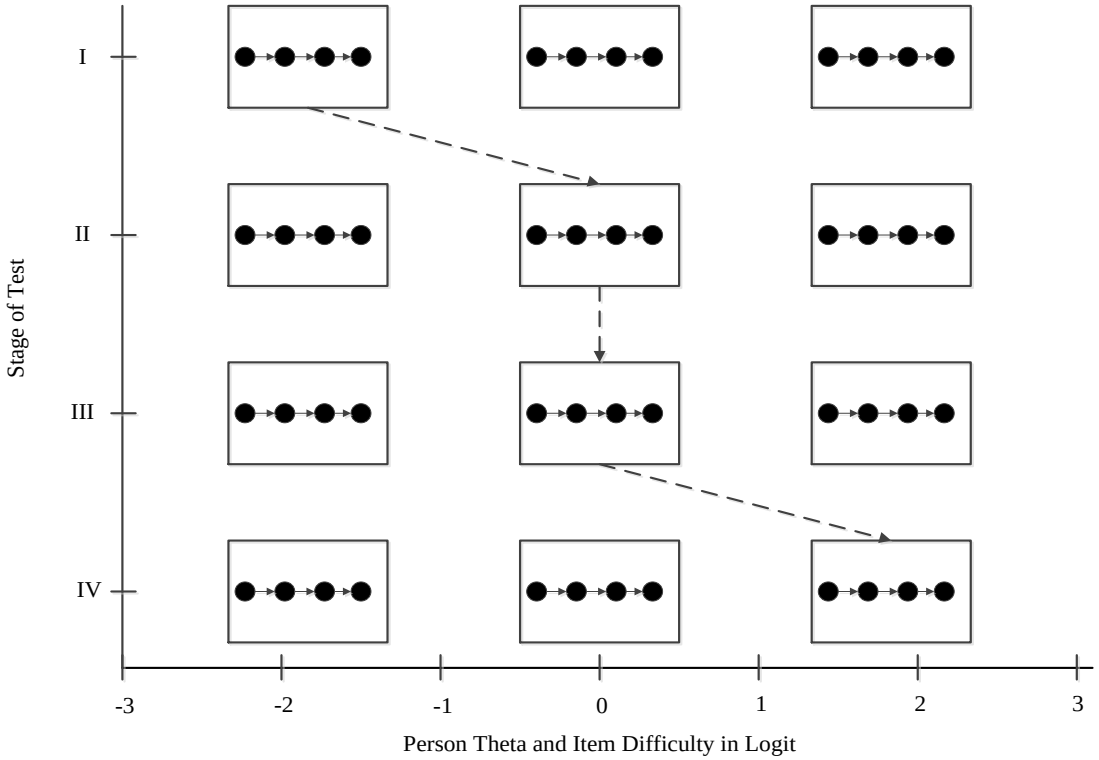

Figure 2. ca-MST Design

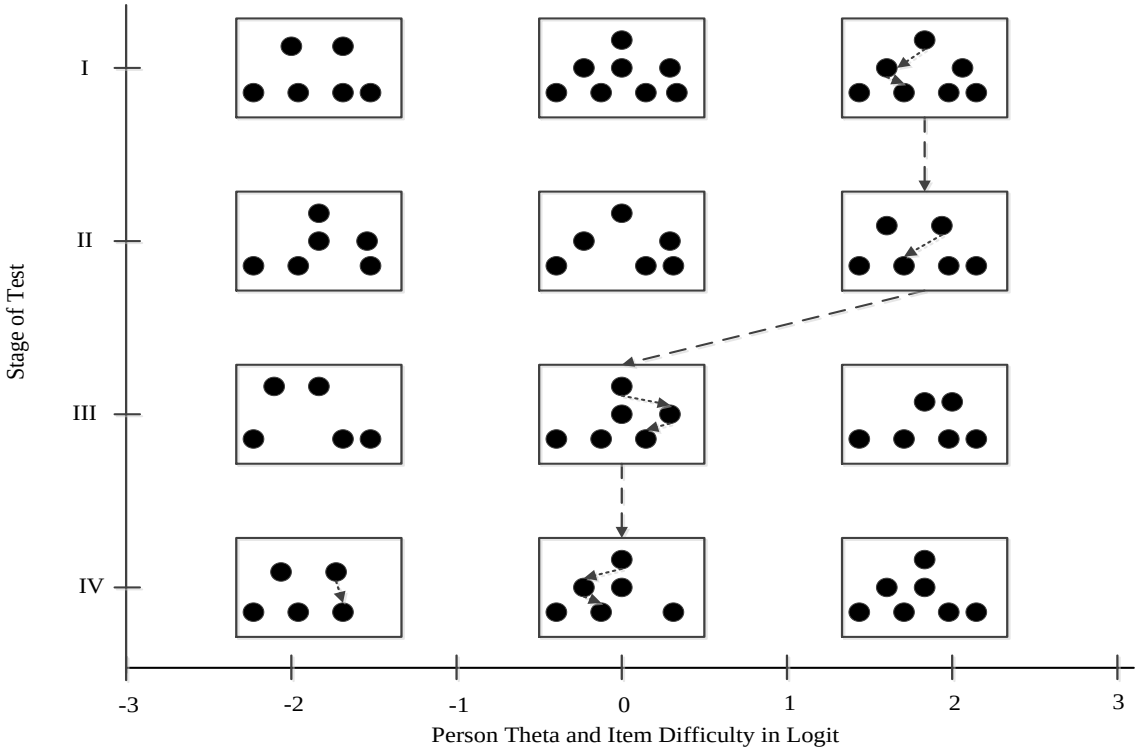

Figure 3. ca-MST (i-CAT) Design

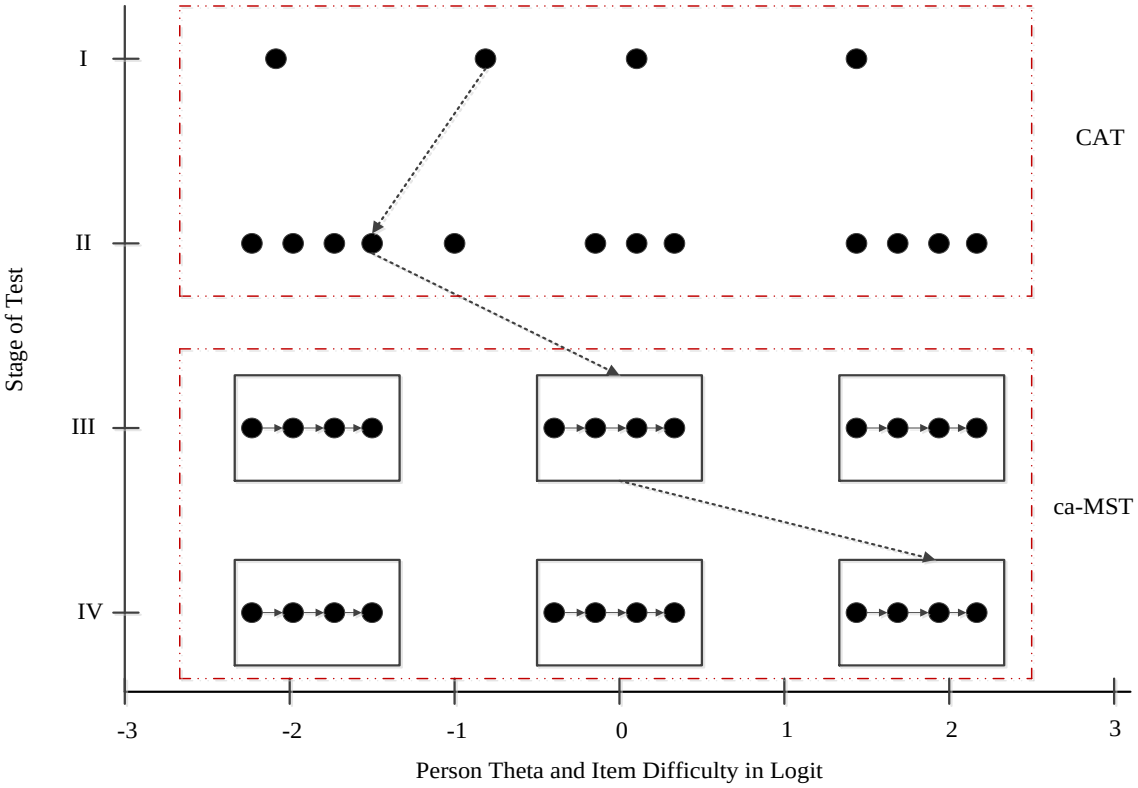

Figure 4. CAT-ca-MST Design

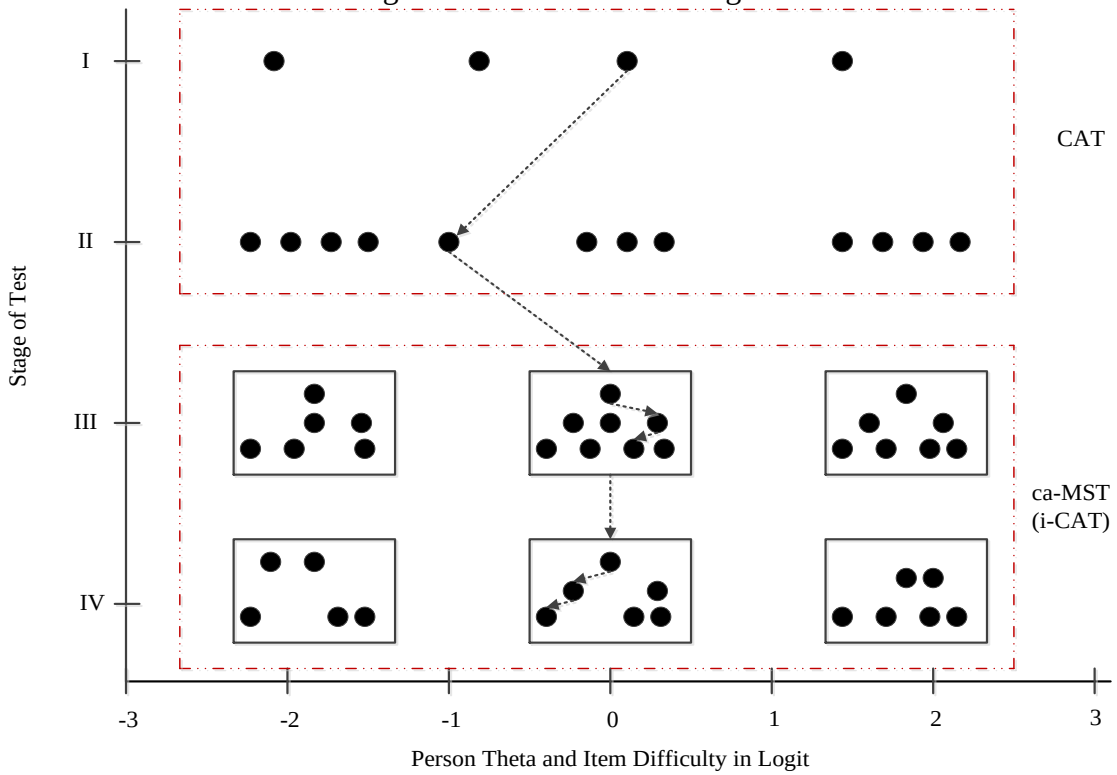

Figure 5. CAT-ca-MST(i-CAT) Design

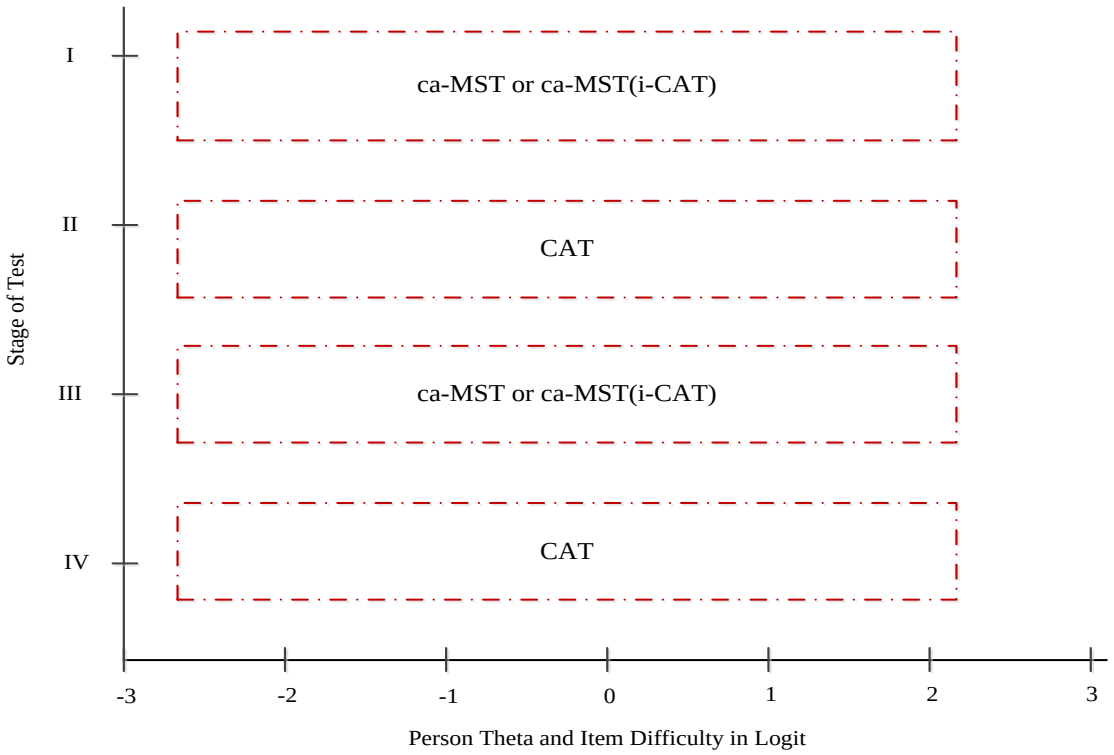

Figure 6. Alternative Design(s)

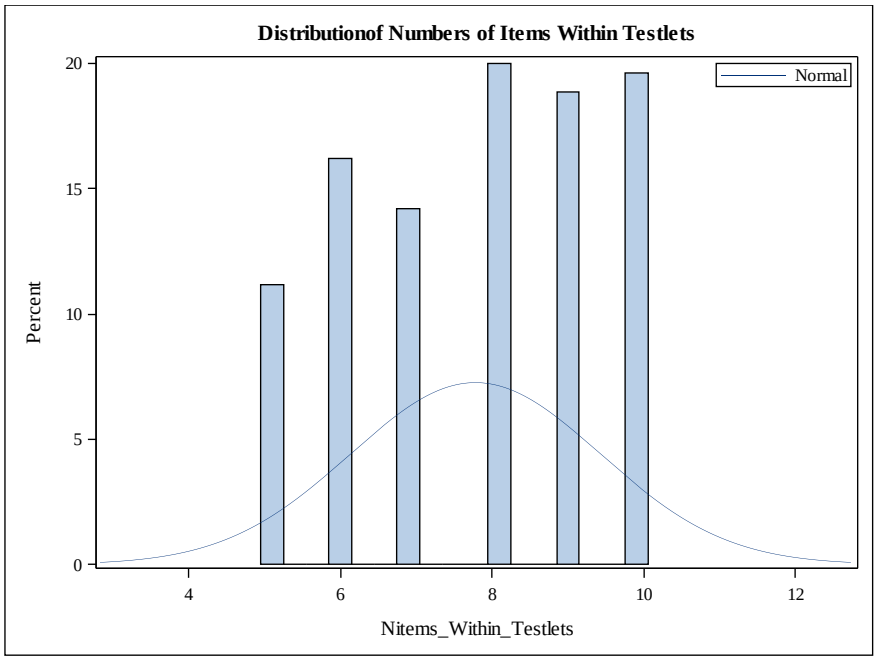

Figure 7. Distribution of Numbers of Items Within Testlets

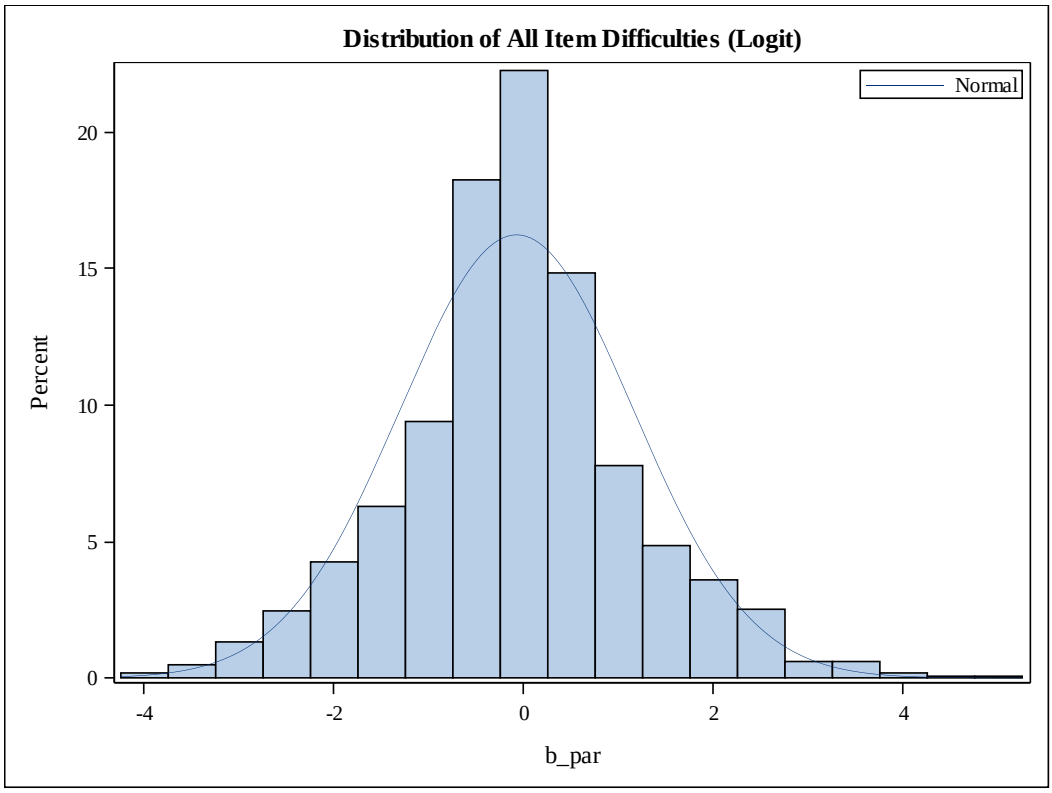

Figure 8. Distribution of Item Difficulties in Bank

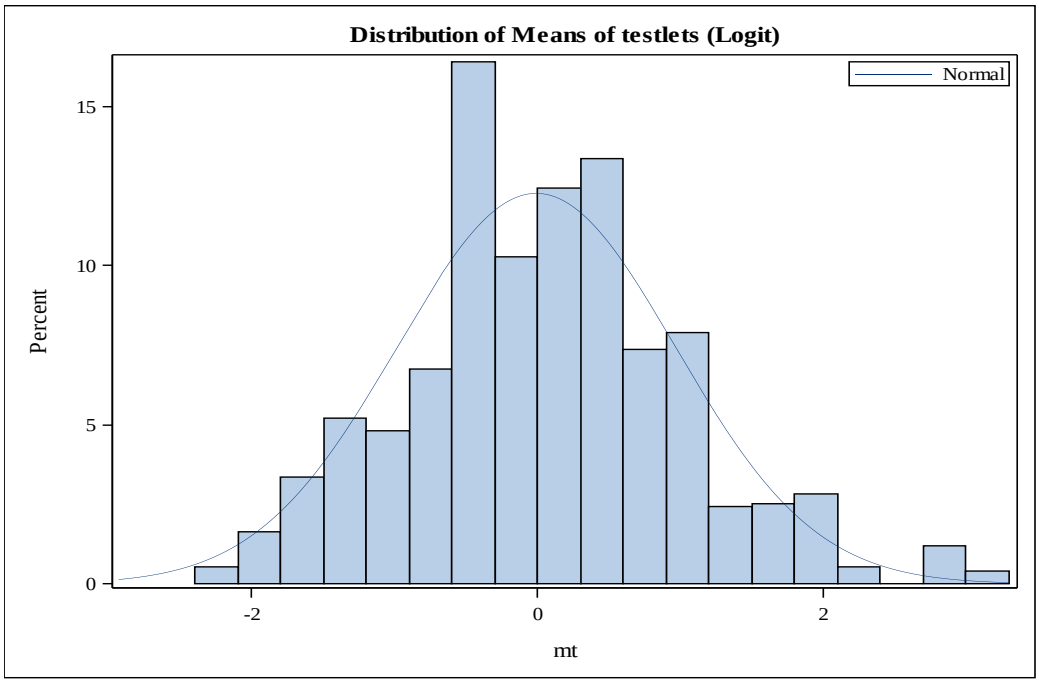

Figure 9. Distribution of Means of Testlets in Bank

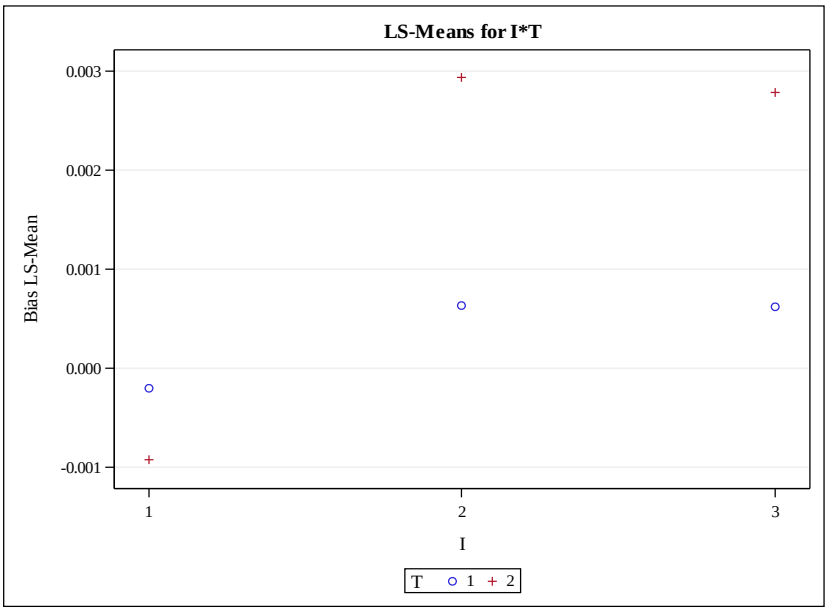

Figure 10. Least Squares Means of Bias for IT(3x2=6) Effect

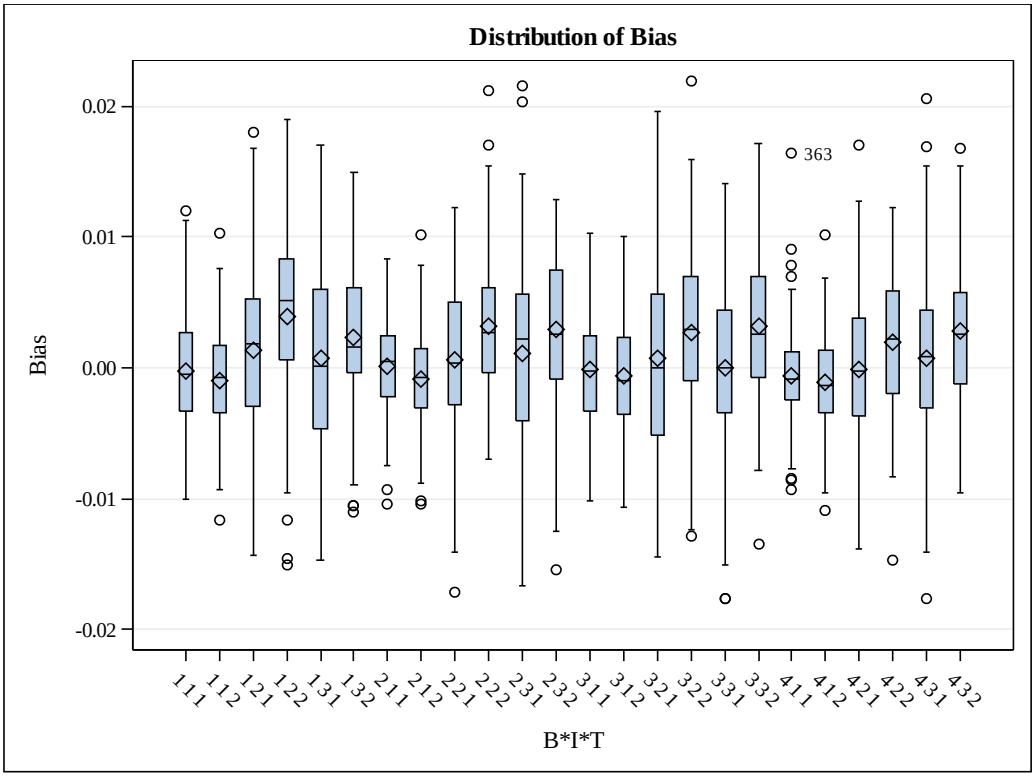

Figure 11. Distribution of Mean Bias for BIT (4 x3x3 = 24) Effect

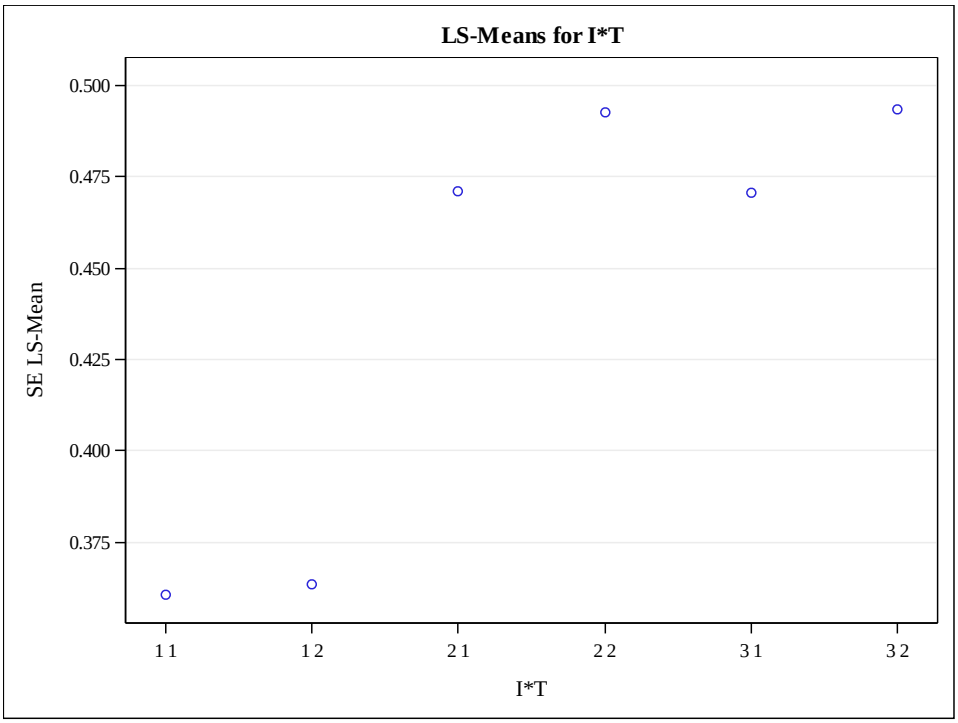

Figure 12. Least Squares Means of SE for IT(3x2=6) Effect

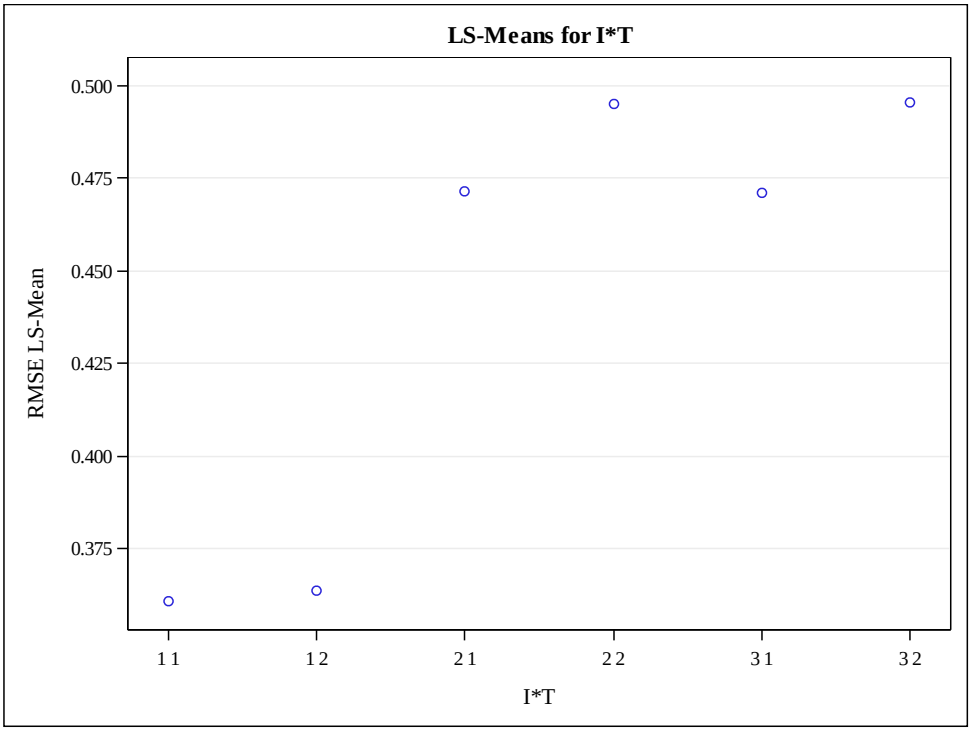

Figure 13. Least Squares Means of RMSE for IT(3x2=6) Effect

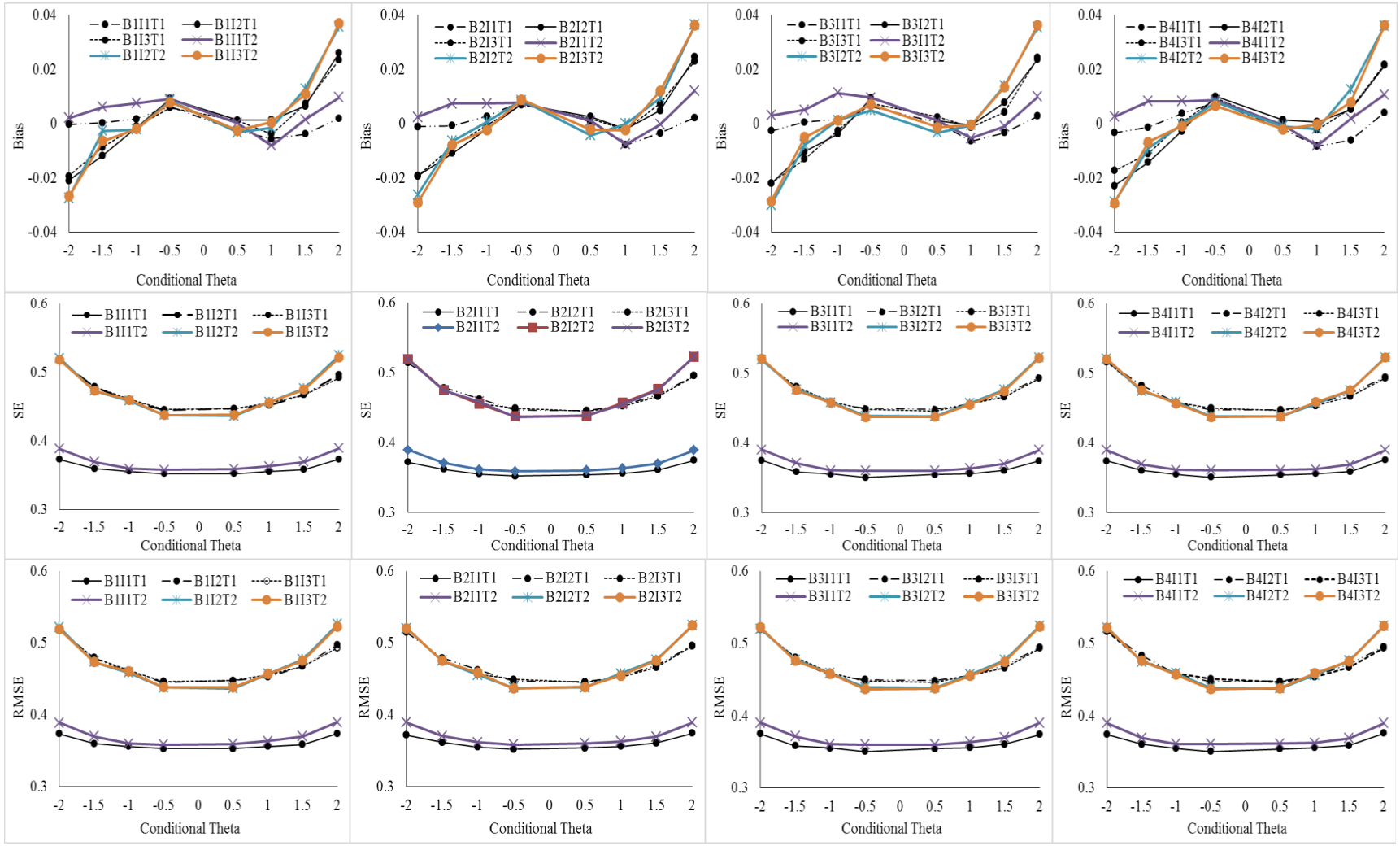

Figure 14. Distributions of Bias, SE, and RMSE for Different B, I, and T

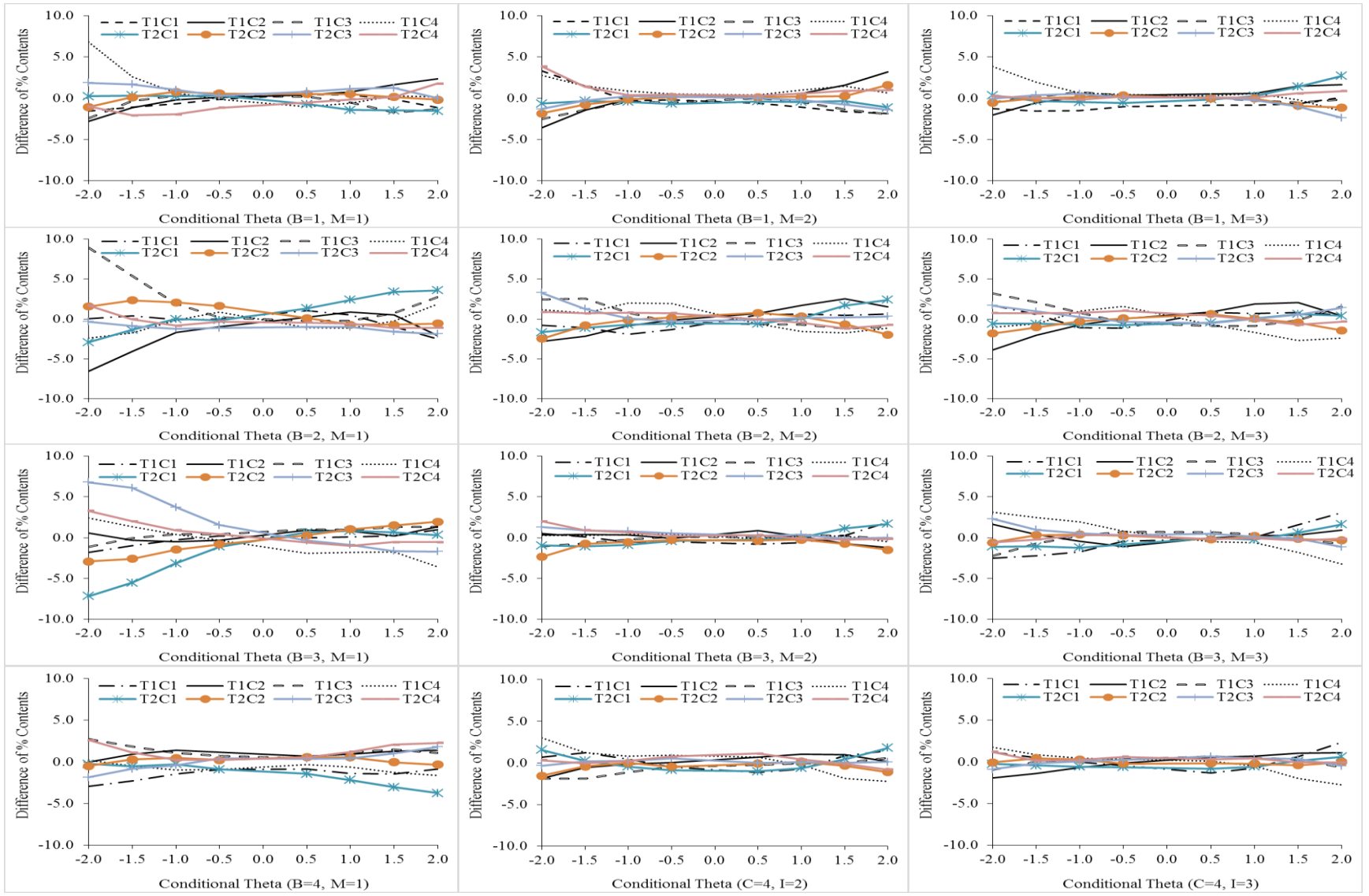

Figure 15. Distributions of Difference of Sub-contents Coverage between Mean (over Replications) and Targeted Percentages

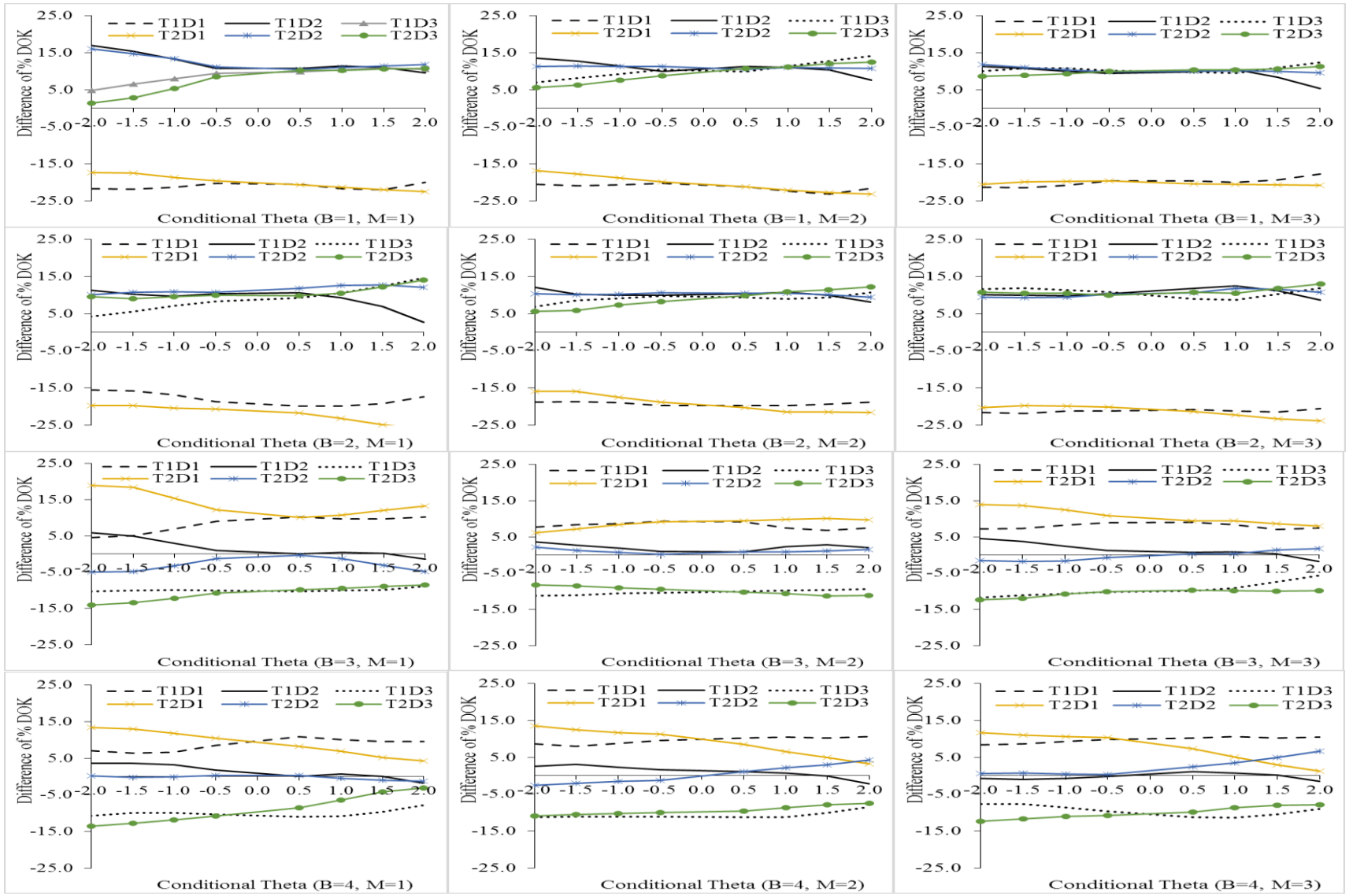

Figure 16. Distributions of Difference of DOK Coverage between Mean (over Replications) and Targeted Percentages

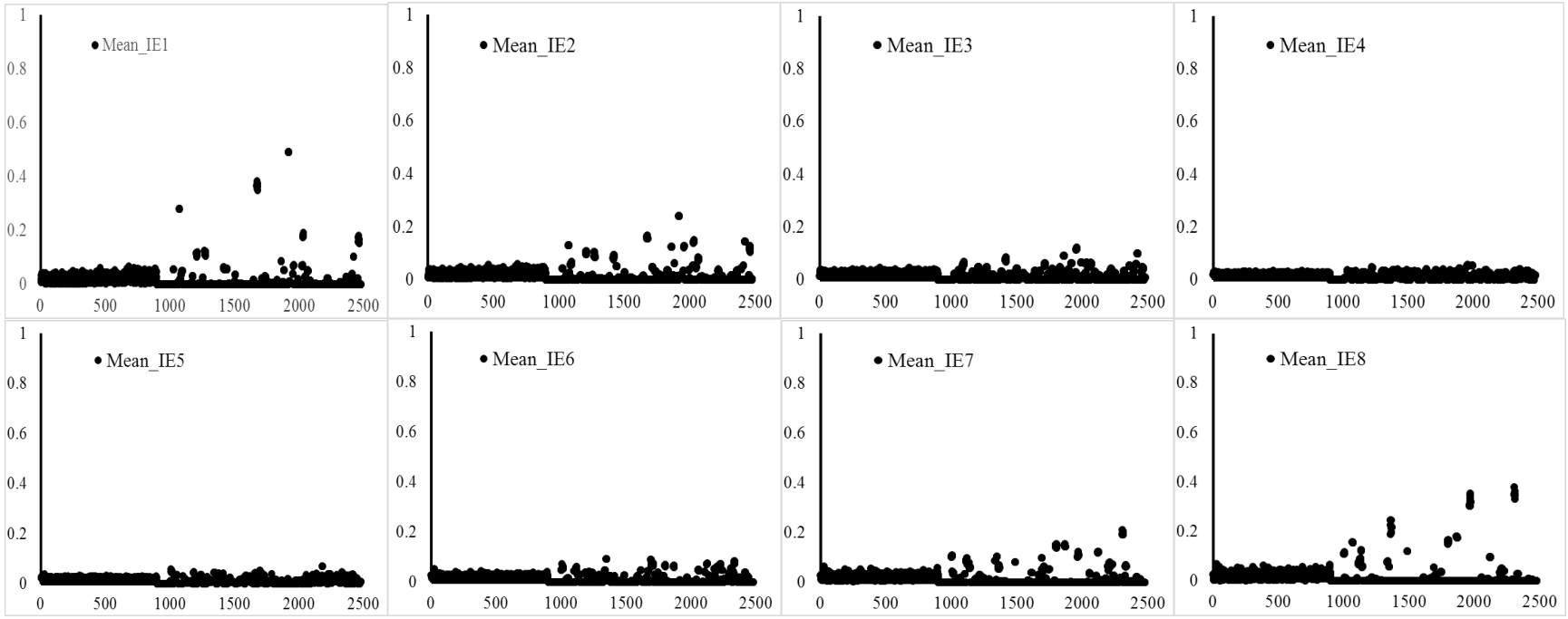

Figure 17. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=1, T=1)

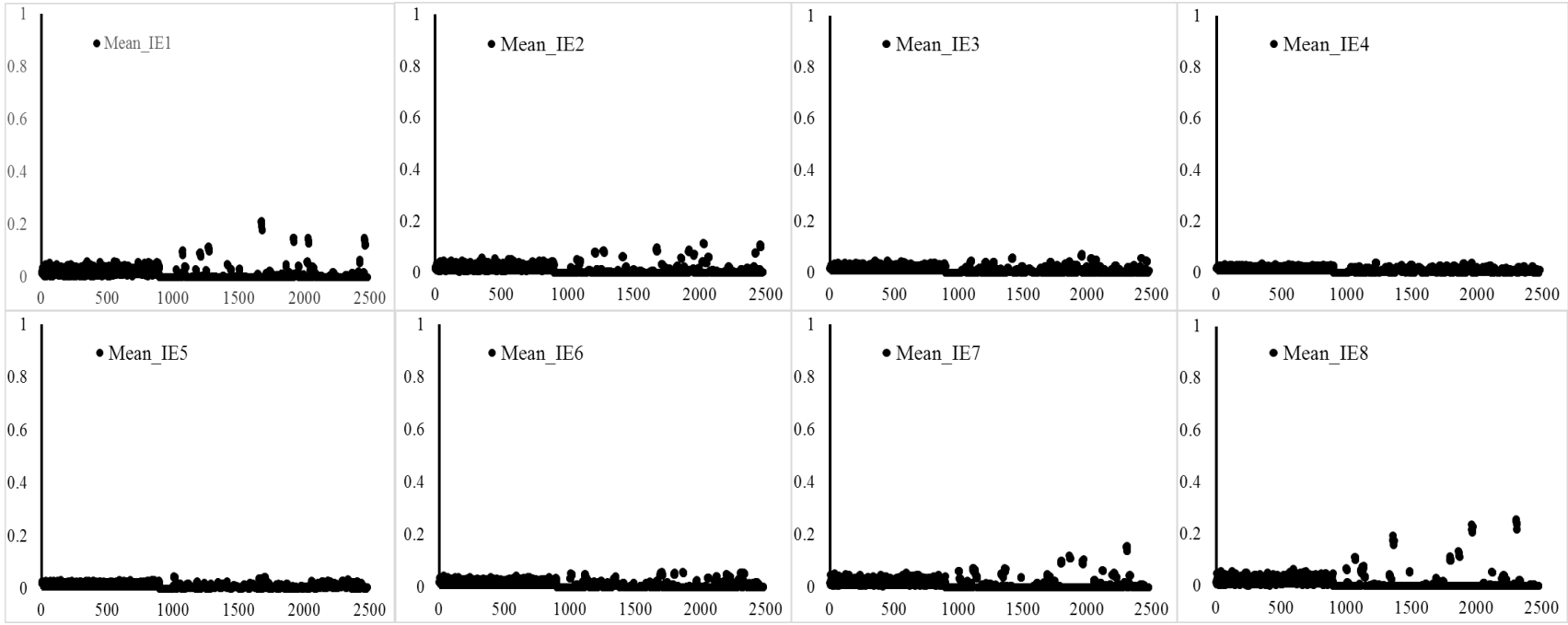

Figure 18. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=2, T=1)

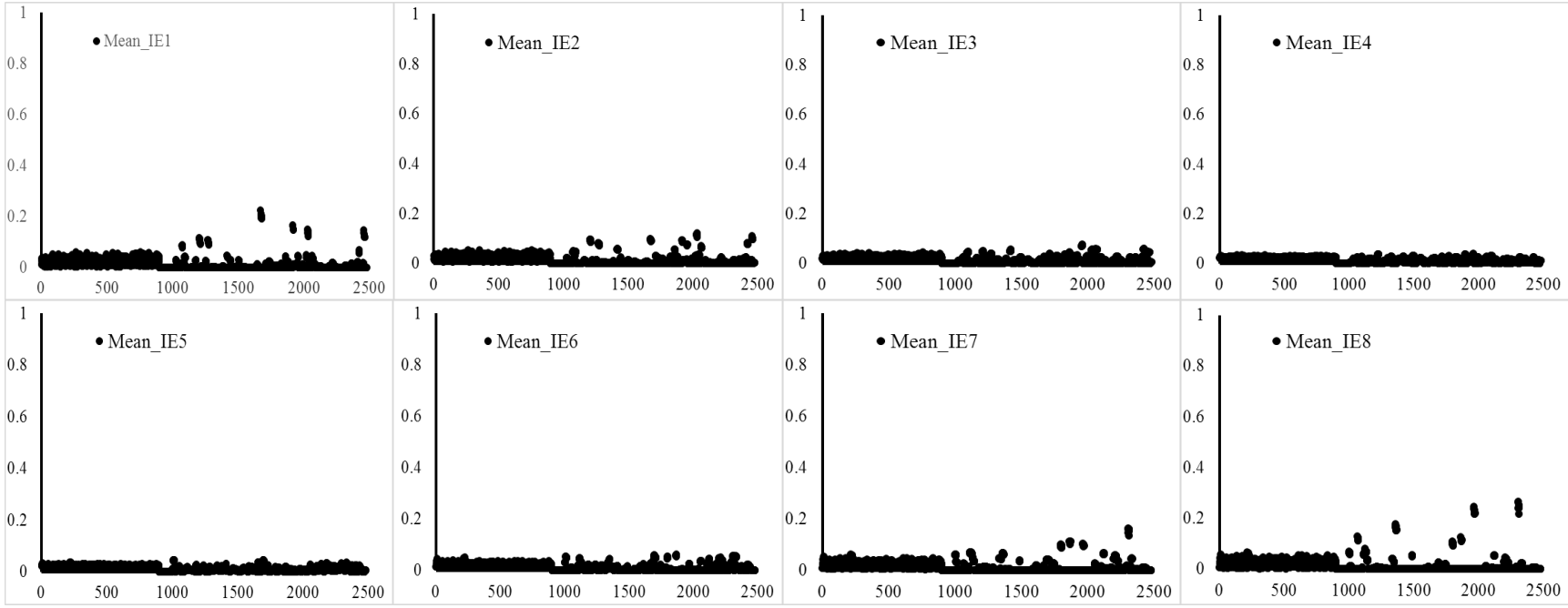

Figure 19. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=3, T=1)

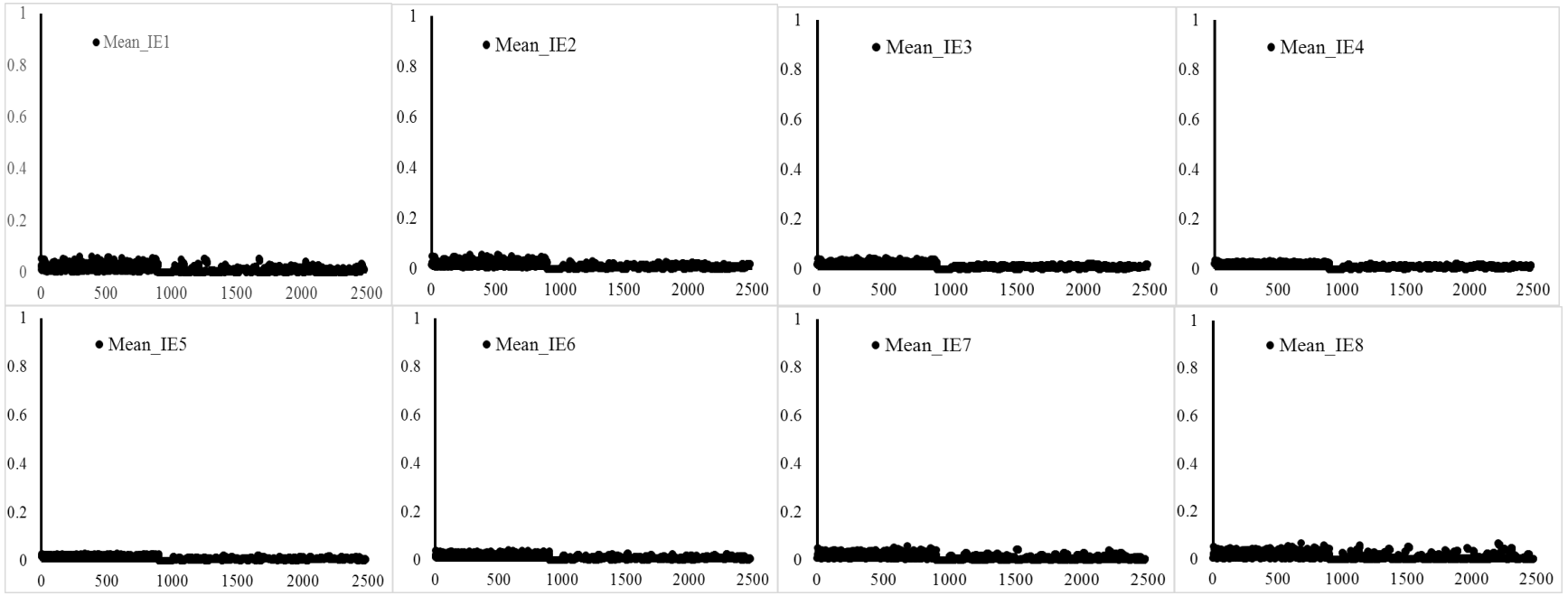

Figure 20. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=1, T=2)

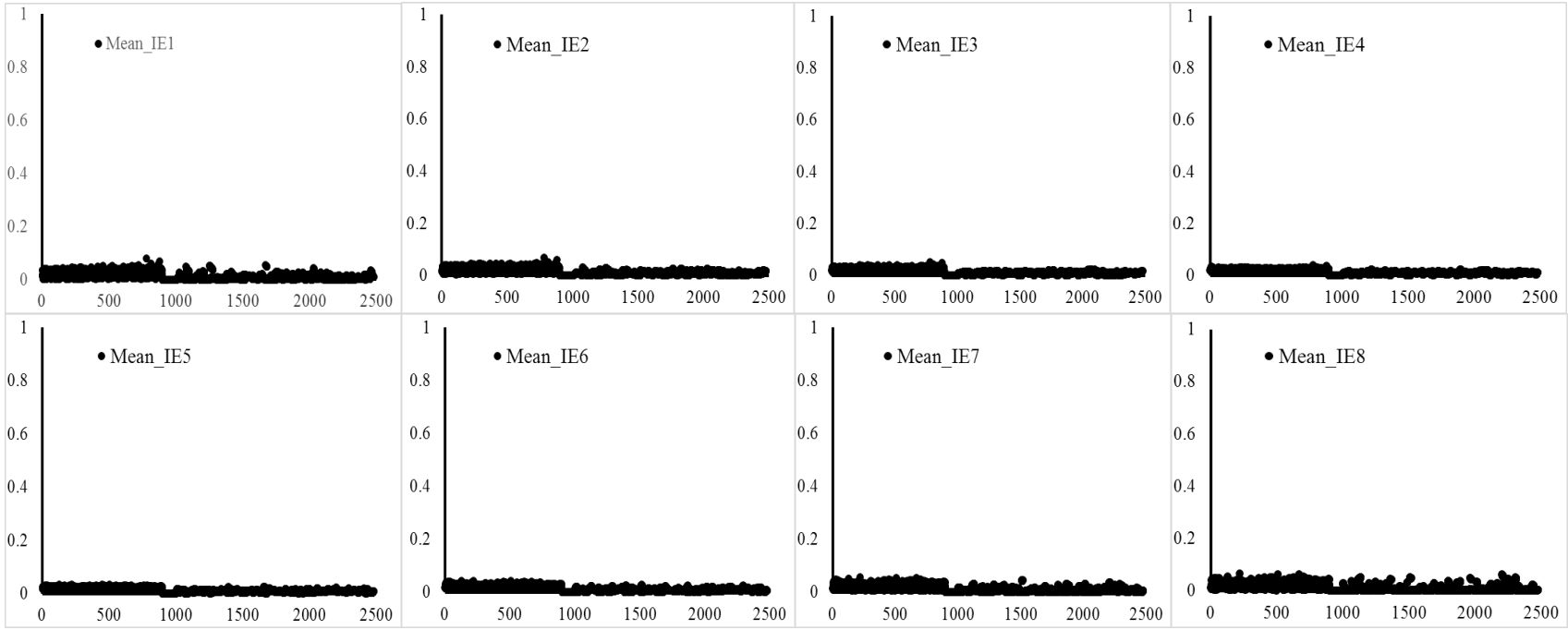

Figure 21. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=2, T=2)

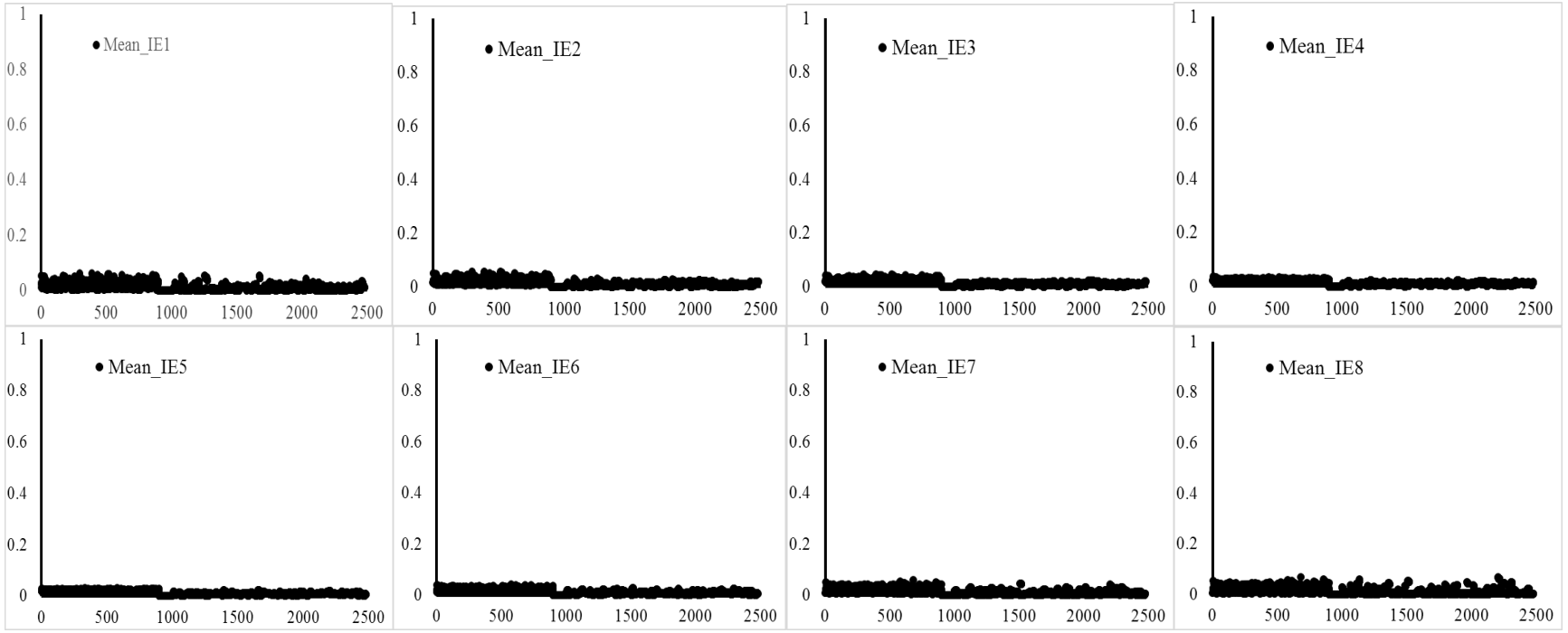

Figure 22. Average Conditional Item Exposure Rate at Eight True Theta (B=1, M=3, T=2)

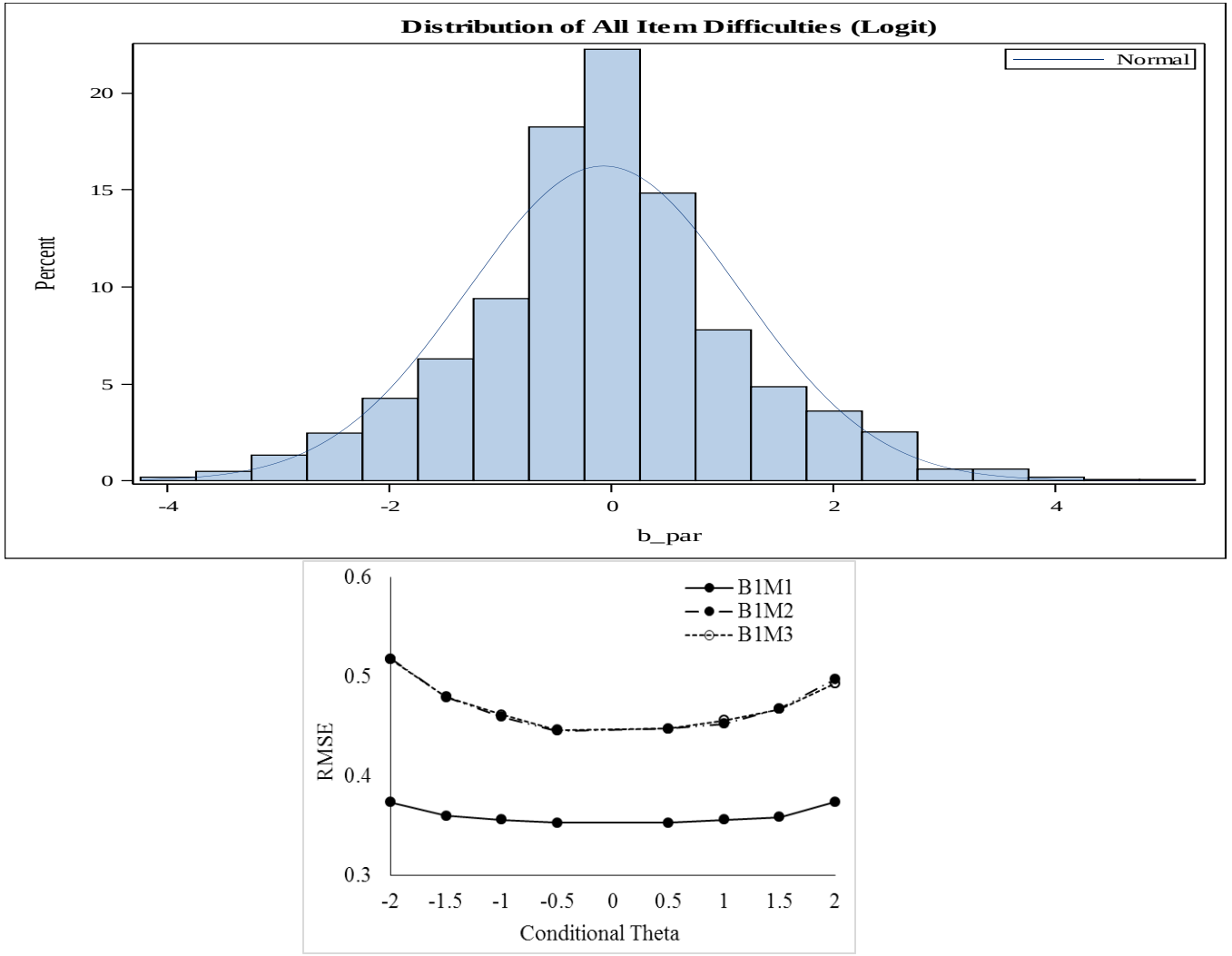

Figure 23. Relationship between Distribution of Item Difficulty in Bank and Test Accuracy