

Running head: ADAPTIVE TESTLETS

A Comparison of Item and Testlet Selection Procedures
in Computerized Adaptive Testing

Leslie Keng

Pearson

Tsung-Han Ho

The University of Texas at Austin

Tzu-An Ann Chen

The University of Texas at Austin

Barbara G. Dodd

The University of Texas at Austin

Correspondence concerning this article should be addressed to Leslie Keng, Pearson, 400
Center Ridge Drive, Austin, TX 78753, E-mail: Leslie.Keng@Pearson.com .

Abstract

Testlet response theory (TRT) is a measurement model that can capture local dependency in testlet-based tests. One of the purported advantages of TRT over the more commonly-used polytomous IRT approach to modeling testlet-based tests is that it allows for *ad hoc testlet construction* in a testlet-based computer adaptive test (CAT). The goal of this study was to investigate the merits of such a CAT design. Specifically, it examined the use of testlet-based CATs that not only chose each *testlet* adaptively, but they also adaptively selected each *item* within the testlet, based on the estimated examinee proficiency. This design was termed a *CAT with adaptive testlets*, and it was compared against a CAT whose within-testlet items were all pre-determined and fixed (termed a *CAT with fixed testlets*). Real data from a large-scale assessment were calibrated using the 3PL-TRT model and used in this simulation study, which compared these testlet-based CAT designs on their measurement and exposure control properties. The study found that the use of adaptive testlets improved measurement precision while achieving better pool utilization rates. The use of an item-level exposure control procedure within a CAT with adaptive testlets resulted in similar measurement precision, but only a modest gain in pool utilization rates, when compared to a CAT with adaptive testlets and no item-level exposure control. This study represented an initial examination of the properties of CATs with adaptive testlets. As such, suggestions for future research are also provided.

A Comparison of Item and Testlet Selection Procedures in Computerized Adaptive Testing

Computer-based testing (CBT) has become a popular alternative mode of test administration to traditional paper-and-pencil (P&P) testing. CBT leverages the benefits of computer technology, leading to several advantages for examinees and test administrators, such as flexibility in scheduling, increased testing opportunities, automated data collection, and prompt score reporting (Bergstorm & Lunz, 1999). It also makes possible the administration of assessments in ways other than the traditional linear fixed format, where every examinee receives the same set of items. One such alternative CBT design is a *computer adaptive test*.

Computer Adaptive Testing

The basic logic behind a computer adaptive test (CAT) is to emulate the test-giving approach of an intelligent human test administrator that adaptively gives test items to each examinee based on his or her evaluation of the examinee's proficiency. This evaluation is typically based on how the examinee has performed so far on the items given by the administrator. If an examinee struggles, then the administrator would give the examinee a low proficiency estimate and administer easier items; if an examinee excels, then the proficiency estimate for the examinee would be high and the administrator would give harder items. Similarly, a CAT algorithm is designed so that examinees receive a tailored test with a set of items that is most accessible to them, based on their estimated proficiency level. In other words, the goals of a CAT is to administer, at each point in the test, items provide maximum information at the examinee's estimated proficiency. Consequently, a CAT is typically shorter than a traditional linear fixed format test while still achieving equivalent or better measurement precision of the examinee's ability (Weiss, 1982; Wainer, 2000). This advantage has made CAT a popular mode of administration. Several large-scale assessments in the educational and

licensure and credential fields, such as the Armed Services Vocational Aptitude Battery (ASVAB), the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), and the National Council of State Boards of Nursing, now offer computer adaptive versions of their assessments (Chang, 2004).

Exposure Control Procedures

One concern that arises from the flexibility of a CAT is that of item exposure and pool utilization. Specifically, items providing maximum information near the center of the ability distribution can be administered too often; while those with maximum information at the extremes are scarcely administered. Uneven item exposure rates, when combined with an item pool that is not sufficiently large, can result in examinees receiving CATs with significant overlap in test items. If examinees can share information with one another between test administrations, then this could pose a large threat to test security. A number of CAT security breach incidents have occurred in recent years and have led to the suspension of CAT administration of a large testing program (Davey & Nering, 2002; Chang, 2004). In addition, poor pool utilization wastes a substantial proportion of items developed for the test. This is an economic concern for testing programs because of the cost and amount of resources that go into the item development process.

Exposure control procedures are often implemented as part of the CAT testing algorithm to address the issues of uneven item exposure and poor pool utilization. A number of CAT exposure control procedures are been developed in literature and implemented in practice. They can be classified into three general categories: randomization procedures, conditional procedures and stratification procedures (Way, 1998).

Randomization procedures control the selection of items by randomly choosing the next item from a set of near-optimal items instead of always choosing the most informative one. Example of randomization exposure procedures include the 5-4-3-2-1 technique (McBride & Martin, 1983; Hetter & Sympon, 1997), the randomesque method (Kingsbury & Zara, 1989), the within-0.10 logits method (Lunz & Stahl, 1998), and the modified within-0.10 logits method (Davis & Dodd, 2003). The advantage of randomization procedures is their relative ease in implementation. However, the maximum exposure rate of items provided by these procedures is not guaranteed and cannot be constrained to a given level (Davis, 2004).

Conditional item exposure procedures control the probability that each item is administered at each step in the test by conditioning it on a given criterion. The criterion is typically based on exposure control parameters, which limit the maximum exposure rate of each item to a predetermined level. Examples of conditional exposure control strategies include the Sympon-Hetter procedure (Sympon & Hetter, 1985), the conditional Sympon-Hetter procedure (Stocking & Lewis, 1998), the Davey-Parshall procedure (Davey & Parshall, 1995), the Stocking and Lewis multinomial procedure (Stocking & Lewis, 1995), and the tri-conditional procedure (Parshall, Hogarty & Kromrey, 1999). The advantage of conditional procedures is that they can guarantee a maximum exposure rate through the exposure control parameters. However, these procedures are generally more difficult to understand and explain. Implementing these procedures also often require complex time-consuming simulations to determine the exposure control parameters. These simulations must be conducted to the operational use of the CAT and can increase implementation complexity (Davis & Dodd, 2003).

Stratification procedures partition the item pool is into strata based on every item's discrimination (a) parameter, and the strata are sorted from low to high discriminating power.

The CAT is then divided into stages that match the number of strata. A pre-determined number of items are administered from each stratum in the corresponding stage of the test. Examples of stratification procedures include the a -stratified design (Chang & Ying, 1996), the a -stratified design with b -blocking (Chang, Qian & Ying, 1999), a multiple-stratification variant of the a -stratified design that incorporate content balancing (Yi & Chang, 2000), and the enhanced a -stratified design (Leung, Chang & Hau, 1999). The rationale behind stratification procedures is to use the least discriminating items at beginning of a CAT, when the accuracy of the proficiency estimate is relatively low, as range finders. The more highly peaked informative items are then saved for the latter stages to pin-point the examinee's ability (Chang & Ying, 1996; Davey & Nering, 2002). Forcing the items with low a -value to be administered at the beginning of a CAT can lead to more even utilization of pool items across a -values, thereby controlling exposure rates within the entire item pool (Chang, 2004).

The Progressive Restrictive Procedure

Because the progressive restrictive procedure (Revuelta & Ponsoda, 1998) was the exposure control method used in the CAT conditions in this study, further details are given here about this procedure.

The progressive restrictive procedure includes aspects of the two other exposure control methods previously proposed by Revuelta and Ponsoda. The *progressive* method (Revuelta & Ponsoda, 1996) is a randomization procedure that extends the unconstrained MI item selection by including a random component that affected the likelihood of an item being chosen. Every unused item in the pool is assigned a weight, defined as a linear combination of the random component and the information provided by the item. The unused item with the highest weight is chosen for administration at each point in the CAT. The weights are computed such that the

random component contributes more to the weight at the start of the test, when less is known about the examinee's proficiency; while item information contributes more to the weight later in the test, when examinee proficiency needs to be pinpointed. The *restrictive* maximum information method (Revuelta and Ponsoda, 1996) is a conditional procedure that, based on the pre-specified maximum exposure rate (k), does not allow any item to be administered to more than $100k\%$ of the examinees.

The *progressive restrictive* procedure combines the two methods by including the computation of weights, as defined in the progressive method, for each unused items, while enforcing the maximum exposure rate specified in the restrictive maximum information method. Therefore, it is a hybrid procedure with both conditional and randomization components. Revuelta and Ponsoda (1998) show that the progressive restrictive procedure produces good measurement precision when the exposure control parameter, k , is sufficiently high (e.g., $k = .40$). The procedure also consistently utilizes all items in the pool while successfully maintaining the maximum exposure rate.

Item Response Theory and Testlets

The implementation of a CAT design is possible because of the development and advancement of a modern test theory, known as *item response theory* (IRT; Rasch, 1960; Lord & Novick, 1968). IRT describes mathematically, the relationship between an examinee's ability and the probability of a given response to a test item based on the item's characteristics. It overcomes many of the shortcomings of classical true score theory (Gulliksen, 1950) because it puts item characteristics and examinee ability on the same scale, thus allowing each examinee's proficiency to be related to his or her item-level performance instead of only the overall test score. This is an important attribute of IRT because it allows different sets of items to be given

to different examinees while still being able to estimate their abilities on the same scale (Embretson & Reise, 2000). This attribute enables the creation of CAT algorithms that build individualized tests for examinees.

IRT does, however, have its limitations. One such limitation stems from the inclusion of an increasingly popular item format called a *testlet*. Wainer and Kiely (1987) first introduced the testlet terminology and defined it as a group of items related to a single content area, developed as a unit, and contains a fixed number of predetermined paths that an examinee may follow. Examples of testlets include a set of reading items associated with a common passage, a group of social studies items referring to a map, or a collection of math items based a graph or table. Testlets are an attractive item format for test developers because of their efficiency in both item development and administration. The use of testlets, however, poses a challenge to use of IRT because of the fundamental assumption of local independence in IRT. Local independence means that, conditional on the examinee's ability, the probability of responding to an item is statistically independent of the probability of responding to any other item (Hambleton & Swaminathan, 1985). Item responses within a testlet are not locally independent because they are related through the common stimulus. As such, using IRT to measure a testlet-based test can therefore lead to inaccurate estimation of examinee and item parameters and overestimation of the precision of these parameters (Tuerlinckx & De Boeck 2001; Sireci, Thissen & Wainer 1991).

Several approaches have been suggested to address the issue of local dependency in testlet-based tests. One viable approach is to define the *testlet* as the unit of measurement and then apply one of the polytomous IRT models (Wainer & Lewis, 1990). Under this approach, a testlet is viewed as a single polytomous item with possible scores ranging from zero to the total number of items in the testlet. This approach has been shown to work well in an array of

situations (Wainer, 1995), including *testlet-based CATs*. Several CAT studies (e.g. Pastor, Dodd & Chang, 2002; Davis, Pastor, Dodd, Chiang & Fitzpatrick, 2003; Davis & Dodd, 2003; Boyd, 2003; Davis, 2004; Davis & Dodd, 2008) have examined the measurement and exposure control properties of testlet-based CATs modeled using a polytomous IRT model, such as the partial credit model (Masters, 1982) or the generalized partial credit model (Muraki, 1992). These studies found that the polytomous IRT approach generally performs well in modeling testlet-based CATs, especially when implemented with one of the randomization exposure control procedures.

The polytomous IRT approach to modeling testlets-based tests, however, has at least two scenarios in which it falls short. The first scenario is when we need to extract more information for each examinee's item response patterns within the testlet. Under the polytomous IRT approach, response patterns that lead to the same total score for a testlet cannot be distinguished, thus decreasing the information gained from the examinee's item responses. Knowing exactly which items the examinee answered correctly and incorrectly could prove beneficial, especially if the items varied by cognitive type or content (Wainer, Bradlow, & Du, 2000).

The second scenario that the polytomous IRT approach cannot be used is in the specific context of a testlet-based CAT where *ad hoc testlet construction* is desired (Wainer, Bradlow & Wang, 2007). The nature of a CAT is to allow items to be selected adaptively. With polytomously-scored testlets, however, there cannot be an interchange of items within a testlet and the items associated with each testlet must be fixed for all examinees. Ad hoc testlet construction allows items administered with a given testlet to be chosen adaptively and can hence vary from one examinee to another. Such a method of testlet administration is more consistent with the adaptive nature of a CAT, and should provide better measurement precision

of examinee proficiency. It should also increase the number of times a stimulus may be used across examinees, leading to more even item exposure and pool utilization. A more recently proposed alternative approach to modeling testlets-based tests that can handle the two above scenarios is *testlet response theory* (TRT).

Testlet Response Theory

In TRT, the *item* remains the unit of measurement. An additional person-specific random effect parameter is included with the most frequently used dichotomous IRT models to account for the shared variance among items within a testlet. This parameter is called the *testlet effect parameter* and is denoted $\gamma_{jd(i)}$ (Bradlow, Wainer & Wang, 1999). As with dichotomous IRT models, TRT models can have as many as three item parameters: difficulty (b_i), discrimination (a_i), and pseudo-guessing (c_i). They also have two person-specific parameters: ability (θ_j), and testlet effect ($\gamma_{jd(i)}$). Below is the function representing the three-parameter logistic testlet response theory model (3PL-TRT; Wainer, Bradlow & Du, 2000),

$$P_{ij}(x_i = 1 | \theta_j) = c_i + (1 - c_i) \left[\frac{\exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))} \right] \quad (1)$$

Other parameterizations of TRT have been proposed, such as the 2PL-TRT (Bradlow, Wainer & Wang, 1999) and a generalized TRT model that can be applied to testlets that include both dichotomous and polytomous items (Wang, Bradlow & Wainer, 2002).

TRT overcomes the two shortcomings of the polytomous IRT approach to modeling testlet-based tests. That is, because item is the unit of measurement, TRT models can extract more information for each examinee's item response patterns within the testlet and allow for ad hoc testlet construction in a testlet-based CAT. However, because of its more recent development, research involving TRT is still scarce and the properties of TRT remain relatively

unknown. Our literature review also found no instances of TRT being implemented in practice in testlet-based tests.

Boyd (2003) was one of the first studies to investigate the measurement and exposure control properties of testlet-based CATs modeling using TRT. Her simulation study compared seven exposure control procedures in two types of testlet-based CAT systems: one was based on the partial credit model (Masters, 1982); while the other was measured using the 3PL-TRT (Wainer, Bradlow & Du, 2000). The study found that the two types of testlet-based CAT systems performed similarly in terms of measurement precision, exposure rates and pool utilization rates. In both cases, the progressive restrictive exposure control procedure (Revuelta & Ponsoda, 1998) with a maximum exposure rate set at 0.30 yielded the best results in both measurement and exposure control properties.

The 3PL-TRT CAT conditions in Boyd's (2003) study were able to capture the first advantage of TRT over polytomous IRT models in extracting information from each examinee's item-level response patterns within testlets. However, the testlets used in both testlet-based CAT systems had pre-determined items associated with them. In other words, once a testlet was chosen for administration, the items given within the testlet were always the same for every examinee; the within-testlet content was *linear* or *fixed*. Thus, the 3PL-TRT conditions in Boyd's study were unable to capture the second advantage of TRT over polytomous IRT models in allowing for ad hoc testlet construction. Boyd listed this as an area for further research, but to date, no studies have examined the properties of such a testlet-based CAT implementation – namely, the use of testlet-based CATs that are adaptive not only *between* testlets, but also *within* testlets. In this study, this type of testlet-based CAT system is referred to as a CAT with *adaptive testlets*. In a CAT with adaptive testlet, every examinee administered a particular testlet does not

necessarily receive the same set of items associated the testlet, but is adaptively administered testlet items that best match his or her estimated proficiency level, conditional on any exposure control procedures.

Research Questions

This study sought to extend Boyd's (2003) research on testlet-based CAT systems modeled with TRT by including CATs with adaptive testlets. The goal of the study is to investigate the properties of CATs with *adaptive testlets*. These properties will be evaluated and compared against the TRT-measured CATs examined by Boyd (2003); that is, testlet-based CATs whose *within*-testlet content is linear or fixed (but are still adaptive *between* testlets) – such CATs are referred to as CATs with *fixed testlets* in this study. Specifically, the study seeks to address the following research questions,

1. To what extent does the use of *adaptive testlets* in a testlet-based CAT system affect its *measurement precision* relative to that of a CAT with *fixed testlets*?
2. To what extent does the use of *adaptive testlets* in a testlet-based CAT system affect its *exposure pool utilization rates* compared to those of a CAT with *fixed testlets*?
3. What are the effects of implementing an item-level exposure control procedure *within* an adaptive testlet (in addition to controlling for exposure *between* testlets) in terms of both measurement and exposure control properties?

Method

Models Compared

To address the research questions, this study included three testlet-based CAT conditions, each modeled with the 3PL-TRT (Wainer, Bradlow & Du, 2000). The three conditions were,

1. CAT with fixed testlets

2. CAT with adaptive testlets, no item-level exposure control
3. CAT with adaptive testlets, with item-level exposure control

It should be noted that while only the third condition included an *item-level* exposure control procedure, all three conditions include the same *testlet-level* exposure control procedure. In other words, for all three conditions, an exposure control procedure was implemented as part of the CAT algorithm to adaptively choose the next *testlet* to administer. Once a testlet was selected, however, the first condition simply administered the pre-determined number of items associated with the testlet; the second condition adaptively chose the same number of items based on maximum information (MI); while the third condition adaptively selected the number of items based on MI, but conditional on the item-level exposure control procedure. As such, comparing results from the first two conditions helped address research questions #1 and #2, while the latter two conditions CAT addressed research question #3.

The *testlet-level* exposure control procedure used in all three conditions was the progressive restrictive procedure (Revuelta & Ponsoda, 1998) with a maximum testlet exposure rate of 0.30. Boyd (2003) found that this procedure and maximum exposure rate yielded the best results in both measurement precision and exposure control. Thus, the first condition in this study matched the best condition from Boyd's (2003) study and served as a baseline for the other two conditions.

The *item-level* exposure control procedure implemented in the third condition was also the progressive restrictive procedure (Revuelta & Ponsoda, 1998), but with a maximum item exposure rate of 0.20. The progressive restrictive procedure was chosen for the item level to match the procedure implemented at the testlet level. It would be unusual in practice for a testlet-based CAT system to use different exposure control procedures at the testlet and item

levels. Because the maximum exposure rate at the testlet level was already 0.30 and every item can only be given if its associated testlet is chosen for administration, this implied that no item could be administered to more than 30% of the examinees. Thus, for the item-level exposure control procedure to have an effect on item selection, a maximum *item* exposure rate needed to be less than 0.30; and 0.20 was chosen for the third condition.

Item Pool

TRT item parameters estimates were obtained from Boyd's (2003) study for a large-scale high-stakes assessment with an item pool consisting of 176 reading passages (testlets) and a total of 1,210 multiple-choice items. The testlets varied in the number of items per passage and corresponding content areas. The pool consisted of 60% six-item, 18% seven-item, 10% eight-item and 12% ten-item testlets. Additionally, the passages were classified into three content areas. In order to ensure that there is a large enough pool of items for the CAT with adaptive testlets conditions, the number of items associated with each reading passage was doubled. In other words, the expanded item pool had 60% 12-item, 18% 14-item, 10% 16-item, and 12% 20-item testlets. The total number of passages was still 176, but a total of 2,420 dichotomous items were available for the study.

Boyd (2003) had used the SCORIGHT software (Wang, Bradlow & Wainer, 2001) to calibrate the item pool using the 3PL-TRT model (see Equation 1). Thus, for each of the 2,420 items, three item parameters were estimated: difficulty (b_i), discrimination (a_i), guessing (c_i). For each of the 176 testlets, there was also a parameter estimate for the variance of the testlet effect, $\text{Var}(\gamma_{jd(i)})$.

Data Generation

A SAS data generation program (Boyd, 2003) based on the 3PL-TRT model was used to generate examinee item responses using the SCORIGHT-calibrated item parameters. Ten replications of 1,000 examinee ability (θ) parameters were generated from the standard normal distribution. Person-specific testlet effect parameters ($\gamma_{jd(i)}$) were randomly selected from a normal distribution with mean zero and variance equal to $\text{Var}(\gamma_{jd(i)})$ for each of the 176 testlets. Each of the three testlet-based CAT study conditions was then run on the same ten samples of 1,000 examinees.

CAT Simulation

In addition to the exposure control procedures described earlier, the three testlet-based CAT conditions in this study all included the following components.

Content Balancing. The CAT algorithm adaptively selected testlets based on MI, contingent not only on the testlet-level exposure control procedure, but also on content balancing specifications. The Kingsbury and Zara (1989) constrained CAT procedure was used for content balancing based on the three content areas and the number of items associated with each reading passage. The target percentages for the content areas and number of items per passage matched those in the test blueprint of the large-scale admission test on which the item pool is based. The target percentages are summarized in Tables 1 and 2.

Insert Tables 1 and 2

Proficiency estimation. The expected a posteriori (EAP) estimation procedure (Bock & Mislevy, 1982) was used to estimate both the interim and final examinee ability (θ) and person-specific testlet effect parameters ($\gamma_{jd(i)}$). For the two adaptive testlet conditions, interim ability

and testlet effect parameter estimations were performed only between testlet administrations and not between item administrations within a testlet.

Stopping rule. A fixed-length CAT of 42 items, consisting of seven testlets each with six items was administered to every examinee in the simulated samples. For the CAT with fixed testlets condition, six items were randomly chosen from the pool of items associated with each passage in advance and designated as fixtures for the testlet. For the two CAT with adaptive testlets conditions, six items that provided maximal information based on each examinee's estimated proficiency level were adaptively selected from each testlet's available pool of items (which varied from 12 to 20 items). For the third condition, MI selection was also contingent on the within-testlet item-level exposure control procedure.

Data Analyses

The three testlet-based CAT conditions were compared on the following measures of estimation precision and accuracy. All statistics were averaged across the 10 replications for each condition.

- Descriptive statistics of the final proficiency (θ) estimates and standard errors
- Correlation between estimated and known proficiency (θ) values

- Bias, defined as:
$$Bias = \frac{\sum_{i=1}^n (\hat{\theta}_i - \theta_i)}{n}$$

- Root mean squared error (RMSE), defined as:
$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}{n}}$$

- Average absolute difference (AAD), defined as
$$AAD = \frac{\sum_{i=1}^n |\hat{\theta}_i - \theta_i|}{n}$$

To compare exposure control properties of the three conditions, the following statistics were computed. As with the indices for measurement precision, all statistics were averaged across the 10 replications for each condition.

- Descriptive statistics of testlet and item exposure rates and pool utilization rates
- Conditional plots of the mean standard errors at various proficiency (θ) levels
- Testlet and item overlap statistics

Testlet or item overlap is defined as the number of testlets or items shared by two examinees. For this study, the mean, minimum and maximum testlet and item overlap statistics were computed across *all* examinees in each condition. The same overlap statistics were also computed separately for examinees of *similar* and of *different* proficiency (θ) levels. Adopting the definition used by Boyd (2003), similar examinees were defined as examinees whose known θ values differ by one logits or fewer; while different examinees were those whose known θ values differ by more than one logits.

Results

Descriptive Statistics

The descriptive statistics of the proficiency (θ) estimates for the three testlet-based CAT conditions, summarized across ten replications, are given in Tables 3 and 4. Note that in each of these tables (and tables to follow), the CAT with fixed testlets condition is denoted *Fixed*; the CAT with adaptive testlets and no item-level exposure control condition is denoted *Adaptive with MI*; and the CAT with adaptive testlets and item-level exposure control condition is represented as *Adaptive with PR*.

Insert Table 3

Table 3 gives the grand mean, standard error of the mean, minimum mean, and maximum mean for the estimated θ values across the ten replications for each of the three CAT conditions. The grand mean of the known θ values across the ten replications was -0.003, with a standard error of the mean of 0.032, a minimum mean of -0.063 and maximum mean of 0.065. The resulting statistics in Table 3 shows that all three CAT conditions yielded very similar mean θ estimates, which were close to the known mean θ statistics.

Insert Table 4

Table 4 shows the mean, minimum and maximum standard deviations for the estimated θ values across the ten replications for each of the three CAT conditions. The values in the table look very similar for the three CAT conditions. The mean standard deviation for the true θ values across the ten replications was 0.0998, with a minimum of 0.962 and a maximum of 1.037. Thus, it appears from Table 4 that each descriptive statistic (mean, minimum and maximum) for the standard deviation of the estimated θ values was lower than that of the known θ value for all three conditions. This pattern was also observed in Boyd (2003) and was attributed to the EAP estimation method used to estimate θ . It had been found that EAP estimation tended to regress toward the mean of the prior distribution (Kim & Nicewander, 1993; Weiss, 1982) and the TRT model may be more susceptible to this tendency.

Measurement Properties

Insert Table 5

Table 5 gives the mean, minimum, and maximum standard errors of the estimated θ 's across the ten replications for each of the three conditions. The standard error of the estimated θ values reflects the measurement precision of each condition, with lower standard errors indicating higher measurement precision. As such, the values in Table 5 shows that the two

CAT with adaptive testlets conditions yielded higher measurement precision than the CAT with fixed testlets. These results reflect the fact that by allowing items to be selected adaptively within each testlet, better measurement precision is achieved by the testlet-based CAT. However, is this gain in measurement precision consistently observed across the proficiency scale? A plot of the grand mean of standard errors conditional of the known θ values is given in Figure 1.

Insert Figure 1

Figure 1 shows that the typical U-shape for this type of conditional plot was observed for all three CAT conditions, indicating higher estimation precision near the center of the θ distribution and lower precision at the extremes. In addition, the standard errors for the CAT with fixed testlet were consistent higher than those for the two CAT with adaptive testlet conditions across the entire θ scale. This implies that the use of adaptive testlets in a testlet-based CAT consistently yielded higher measurement precision for examinees of all proficiency levels. Note also that in both Table 5 and Figure 1, the two adaptive testlets condition produced very similar results in terms of standard errors. In most cases, the implementation of exposure control procedure in a CAT algorithm decreases measurement precision as items with less than maximal information are being chosen for administration. However, the fact that the *Adaptive Testlet with PR* condition yielded similar measurement precision as the *Adaptive Testlet with MI* condition shows that the within-testlet item-level exposure control enforced by the progressive restrictive procedure did not lead to any substantial decrease in measurement precision. These results are consistent with the good measurement precision found for the progressive restrictive procedure by Revuelta and Ponsoda (1998) as well as by Boyd (2003).

Indicators of measurement accuracy for the three CAT conditions are provided in Table 6.

Insert Table 6

The correlation between known and estimated θ , bias, root mean squared error (RMSE) and average absolute difference (AAD) averaged across the ten replications all showed similar patterns for the three CAT conditions. The correlations were all above 0.90, while the biases were all very close to zero, especially when these value were rounded to the second decimal place. This indicates very good measurement accuracy for all three conditions; with the two CAT with adaptive testlet conditions achieving slightly higher, but not practically significant, accuracy than the CAT with fixed testlets condition. The RMSE and AAD statistics also showed the pattern of measurement accuracy being slightly higher for the two CAT with adaptive testlet conditions. However, the small differences in these statistics between the fixed and adaptive testlet conditions were not practically significant. These finding were further corroborated in Figure 2 by the conditional plot of the difference between known and estimated θ values for the three conditions. The lines for the three conditions were barely distinguishable, especially near the middle of the proficiency distribution. Thus, measurement accuracy was similar and consistently good for the three testlet-based CAT conditions in this study.

Insert Figure 2

Exposure Control

Table 7 gives the descriptive statistics for the mean and standard deviation of the *testlet* exposure rates for the three CAT conditions. The statistics in this table is best examined with the findings in Table 8, which shows the average pool utilization frequency of the 176 testlets across the ten replications for the three CAT conditions.

Insert Tables 7 and 8

The fact that the grand means of the testlet exposure rates (in Table 7) for the three conditions are exactly equal was not an unexpected result given that the ratio of pool size to test

length was the same across conditions (Chen, Ankenmann & Spray, 2003). Furthermore, because the between-testlet CAT algorithm were essentially the same for the three CAT conditions, it should come as no surprise that the exposure control results at the testlet level (in Tables 7 and 8) were very similar across conditions.

It should also be noted that the means of the maximum testlet exposure rates (in Table 7) and maximum utilization frequencies (in Table 8) imply that the testlet-level progressive restricted procedure was generally successful at enforcing the maximum exposure rate of 0.30. In addition, Table 8 shows that the frequencies and percentages of testlets that were never administered were effectively zero. These observations combined with the earlier measurement precision results were further validation of the sound measurement and exposure control properties of the progressive restrictive procedure that had been found in previous studies (Revuelta & Ponsoda, 1998; Boyd, 2003).

Table 9 gives the descriptive statistics for the mean and standard deviation of the *item* exposure rates for the three CAT conditions. The results in Table 9 are best interpreted when complimented with the findings in Table 10, the average pool utilization frequency of the 2,042 items across the ten replications for the three conditions.

Insert Tables 9 and 10

The most notable trend in examining these two tables was the substantially better pool utilization rates of the two adaptive testlets conditions. In Table 10, we see that, on average, over 56% (or 1,367) of the items in the pool were never utilized for the *Fixed* condition. That statistic drops to 32% (or 782) of the items for the *Adaptive with MI* condition, and to 30% (or 727) of the items of the *Adaptive with PR* condition. Thus, by allowing items to be selected adaptively within a testlet, considerably more items in the pool were administered to examinees.

An additional comparison can be made between the two CAT with adaptive testlet conditions. In Table 9, the maximum of the mean item exposure rates for the *Adaptive with PR* condition was 0.201, which shows that the item-level progressive-restricted exposure control procedure successfully enforced the maximum exposure rate of 0.20. This is considerably lower than the 0.30 for the *Adaptive with MI* condition, which had no item-level exposure control and hence maintained the testlet-level maximum exposure rate of 0.30. However, in comparing the distribution of exposure rates in Table 10, we see that the main difference between the two conditions was that items with exposure rates higher than 0.20 in the *Adaptive with MI* condition were “pushed down” to have exposure rates between 0.16 and 0.20 in the *Adaptive with PR* condition. The distributions of items with exposure rates less than 0.16 were very similar for the two conditions. The difference in percentages of items never administered under the two adaptive testlet conditions was also not as substantial, especially when compared to the differences between these two conditions and the *Fixed* condition. Thus, it would appear that the use of the progressive restrictive procedure at the item level within a testlet led to only moderate improvements in the item exposure and utilization rates.

Overlap Statistics

Overlap statistics is an additional measure of exposure control effectiveness as it captures the number of testlets or items two examinees have in common. Tables 11 and 12 provide the *testlet* overlap statistics for the three CAT conditions. Table 11 gives the testlet overlap statistics for all examinees; while Table 12 breaks down the same statistics for examinees of similar and difference proficiency levels. Recall that similar examinees were defined as examinees whose known θ values differ by one logits or fewer; while different examinees had known θ values that differed by more than one logits (Boyd, 2003).

Insert Tables 11 and 12

In comparing the testlet overlap statistics for the conditions in these two tables, we note a similar trend to what was observed earlier for the testlet-level exposure and utilization rates (in Tables 7 and 8). Namely, the overlap statistics for the three conditions were strikingly similar. Overall, the average testlet overlap rates for the three conditions were all around 0.84, implying that, on average, examinees had less than one testlet in common. For different examinees, the testlet overlap rates went down to about 0.62; while for similar examinees, the rates were just over 1 for all conditions. The fact that even similar examinees had on average only one (out of a total of 7) testlets administered in common validated, once again, the effectiveness of the testlet-level progressive restrictive exposure control procedure in all three CAT conditions.

Tables 13 and 14 give the *item* overlap statistics for the three CAT conditions. Table 13 provides item overlap statistics across all examinees; while Table 14 separates the item overlap statistics for similar and difference examinees.

Insert Tables 13 and 14

The item overlap statistics in Tables 13 and 14 may appear to be more different across the three conditions than those at the testlet level. However, one should remember that each CAT had a total test length of 42 items. Thus, a difference in overlap of one or two items (out of 42) between two examinees was likely not of any practical significance. As such, while the *Fixed* condition generally had the highest overlap of items between examinees – similar, different or overall – and the *Adaptive with PR* condition, in general, had the lowest item overlap, the differences in overlap statistics across the CAT conditions were not practically significant.

Discussion

This goal of this study is to investigate the measurement and exposure control properties of CATs with adaptive testlets. The study is an extension of Boyd's (2003) research on testlet-based CAT systems modeled with TRT. The baseline condition in the study (CAT with fixed testlets) is the condition with the best overall results in Boyd's study. Two CAT conditions that implement adaptive testlets – one with within-testlet item-level exposure control and one without – are compared to the baseline condition. The findings from the analysis help answer the three research questions of interest.

Research Questions

1. *To what extent does the use of adaptive testlets in a testlet-based CAT system affect its measurement precision relative to that of a CAT with fixed testlets?*

Comparison of the measurement properties of the CAT conditions shows that the use of adaptive testlets in a testlet-based CAT can improve measurement precision while maintaining similar measurement accuracy as a CAT with fixed testlets. The improvement in measurement precision in CATs with adaptive testlets is most clearly illustrated in Figure 1, where the conditional grand mean standard errors for the two adaptive testlet conditions are distinctly lower than that of the fixed testlets condition across the entire θ scale.

2. *To what extent does the use of adaptive testlets in a testlet-based CAT system affect its exposure rates and pool utilization rates compared to those of a CAT with fixed testlets?*

In comparing the exposure control properties of the three testlet-based CAT conditions in the study, we find that at the *testlet* level – that is, testlet exposure and utilization rates – the three conditions perform similarly well. This should not come as a surprise because the CAT algorithm for adaptively selecting between testlets is the same for the three conditions. The main

differences between the conditions are found at the *item* level. By allowing items to be adaptively chosen within each testlet, the proportion of items in the pool that are never administered decreases considerably compared to when items associate with each testlet are fixed. The overlap statistics for the three conditions, however, are similar. Thus, the use of adaptive testlets does not practically decrease the amount of testlet or item overlap between examinees. Of course, the overlap statistics for the CAT with fixed testlets condition are already quite good. Thus, there is not much room for improvement for the two CAT with adaptive testlet conditions in terms overlap properties.

It should be noted that the statistics observed in this study for the baseline (CAT with fixed testlets) condition mirror those found for the analogous condition in Boyd (2003). Thus, our baseline results also serve as a cross-validation of what Boyd found for that condition. In Boyd's study, this condition represents the best of seven testlet-based CAT conditions measured under TRT in both measurement and exposure control properties. The fact that the CATs with adaptive testlets in this study are able to further improve on the measurement precision and item utilization rates empirically demonstrates the merits of allowing items to be selected adaptively within a testlet, as suggested in literature (Wainer, Bradlow & Du, 2000).

3. *What are the effects of implementing an item-level exposure control procedure within an adaptive testlet (in addition to controlling for exposure between testlets) in terms of both measurement and exposure control properties?*

Comparisons between the two CAT with adaptive testlets conditions show that the implementation of an exposure control procedure at the item level yielded similar measurement precision and accuracy, but only modest improvements in item exposure and utilization rates. Thus, while there does appear to be slight benefits to including a within-testlet item-level

exposure control procedure for CAT with adaptive testlets, the modest gains need to be weighed against the added complexity in implementing and administering the CAT

Limitations and Future Research

The findings in this study demonstrate one of the purported advantages of using TRT over polytomous IRT to model testlet-based tests; namely, the use of *ad hoc testlet construction* in testlet-based CATs. While the results are encouraging for this testlet-based CAT design, they should only be viewed as initial findings in an area of CAT research that is, to date, relatively unexplored. Future studies should seek to cross-validate the results found in this study.

One somewhat unanticipated finding in this study is the relatively small impact on item utilization made by the inclusion a within-testlet item-level exposure control procedure. The authors had anticipated that controlling for exposure at the item level would lead to a substantial decrease in the proportion of items in the pool that are never administered across examinees. That decrease though was rather modest – only a 2% improvement (gain of 55 out of 2042 items) when compared to the adaptive testlet with no item-level exposure control. However, before concluding that an item-level exposure control procedure is not needed within an adaptive testlet, one should consider other factors that may have contributed to this modest gain. Such factors could include the test length, the number of items associated with each testlet, and the number of items administered within each testlet (i.e. testlet length). Future studies can manipulate one or more of these factors to see whether the item-level exposure control procedure makes a larger impact on the item exposure and utilization rates.

Also, recall that in the baseline (CAT with fixed testlets) condition, the six items always administered with a given testlet are pre-determined and are randomly chosen from the pool of items associated with that testlet. This scenario is unrealistic in at least two ways and likely

contributed to the higher percentage of items never administered for this condition. First, in practice, the items associated with a testlet are never randomly chosen. Careful psychometric and content considerations are usually given during test construction in selecting the set of items administered with a common stimulus, such as a reading passage. Second, every testlet in the item pool for this study have more than six items associated with it – they have either 12, 14, 16 or 20 associated items. Because only six items are administered with a testlet, this means that either 6, 8, 10 or 14 items associated with a testlet have no chance of ever being used. This scenario rarely occurs in practice because it would mean that a significant portion of the item pool is, by default, never used. A more realistic scenario is to allow different permutations of the same testlet – that is, different versions of a given testlet based on a common stimulus, but with different subsets of associated items. These different permutations of a testlet could be constructed according to the testing program’s content blueprint and psychometric specifications. Future studies should allow different permutations of testlets in the item pool and see how much the testlet and item pool utilization rates for the CAT with fixed testlets condition improve in comparison to CATs with adaptive testlets.

Lastly, while the use of adaptive testlets appears to lead to gains in measurement precision and pool utilization, these benefits need to be weighed against the non-psychometric considerations in building a testlet. For example, adaptively selecting items to include with a testlet may not appropriately account for context effects or satisfy the content constraints for the testlet. Content balancing procedures, such as Kingsbury and Zara (1989), could be included as part of the within-testlet item selection algorithm. However, including content balancing procedures may negatively affect the measurement precision gained by using adaptive testlets.

Future studies should therefore examine the effect of item-level content balancing procedures on the measurement properties of CATs with adaptive testlets.

Despite these limitations, this study does represent an initial exploration into the use of adaptive testlets in a testlet-based CAT. With the increasingly demand on many testing programs to move to CBT, the lingering security concerns about CBT and CAT, and the growing popularity of the testlet format in large-scale assessments, research around the properties and viability of testlet-based CATs becomes ever more important. Thus, the findings from this study should help inform researchers and practitioners on the utilities of implementing CATs with adaptive testlets and serve as a starting point for future research in this area.

References

- Bergstrom B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Boyd, A. M. (2003). *Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems*. Unpublished dissertation.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153-168.
- Chang, H. H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences*, (pp. 117-133). Sage Publications.
- Chang, H. H., Qian, J., & Ying, Z. (2001). A-stratified multistage CAT with b blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). *Performance of item exposure control methods in computerized adaptive testing: Further explorations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Chen, S., 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.
- Davey T., & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. Mills, M. T. Potenza, J. J. Fremer and W. C. Ward (Eds.), *Computer-Based Testing: Building the Foundation for Future Assessments*, Mahwah, NJ: Lawrence Erlbaum Associates.
- Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, 28(3), 165-185.
- Davis, L. L., & Dodd, B. G. (2003). Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. *Applied Psychological Measurement*, 27(5), 335-356.
- Davis, L. L., & Dodd, B. G. (2008). Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Partial Credit Model. *Journal of Applied Measurement*, 9(1), 1-17.
- Davis, L. L., Pastor, D. A., Dodd, B. G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, 4(1), 24-42.

- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Gulliksen, H. (1950). *Theory of Mental Tests*. Hillsdale, NJ: Lawrence Erlbaum.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. K. Waters, and J. R. McBride (Eds.), *Computerized adaptive testing-from inquiry to operation* (pp. 141-144). Washington, D.C.: American Psychological Association.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359-375.
- Leung, C., Chang, H. H., & Hau, K. (1999). *An enhanced stratified computerized adaptive testing design*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lunz, M. E., & Stahl, J. A. (1998). *Patterns of item exposure using a randomized CAT algorithm*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- 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* (pp.223-226). New York, Academic Press.

- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Parshall, C. G., Hogarty, K. Y., & Kromrey, J. D. (1999). *Item exposure in adaptive tests: an empirical investigation of control strategies*. Paper presented at the annual meeting of the Psychometrics Society, Lawrence, KS.
- Pastor, D.A., Dodd, B.G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, 26 (2), 147-163.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
- Revuelta, J., & Ponsoda, V. (1996). Metodos sencillos para el control de las tasas de exposicion en tests adaptativos informatizados [Simple methods for item exposure control in CATs]. *Psicologica*, 17, 161- 172.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 26(2), 144-163.
- Sireci, S. G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.
- Stocking, M. L., & Lewis, C. (1995). *A new method for controlling item exposure in computer adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, *Journal of Educational and Behavioral Statistics*, 23, 57-75.

- Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163-182). Netherlands: Kluwer Academic Publishers.
- Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.
- Tuerlinckx, F., and De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6, p.181-195.
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Psychological Measurement*, 8 (2), 157-187.
- Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NH: Lawrence Erlbaum Associates.
- Wainer H., & Kiely G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185-201.
- Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 24(3), 185-201.
- Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 245-269). Netherlands: Kluwer Academic Publishers.

- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Wang, X., Bradlow, E. T., & Wainer, H. (2000). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement*, 26(1), 109-128.
- Wang, X., Bradlow, E. T., & Wainer, H. (2001). The SCORIGHT computer program [Computer program]. Princeton, NJ: Educational Testing Service.
- Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17(4), 17-27.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Yi, Q., & Chang, H. H. (2000). *Multiple stratification CAT designs with content control*. Unpublished manuscript.

TABLE 1: Target Percentages for Content Areas used for Content Balancing

Content Area	Target Percentage
1	40%
2	36%
3	23%

TABLE 2: Target Percentages for Number of Items per Passage used for Content Balancing

Number of Items Associated with Passage	Target Percentage
12 Items	42%
14 Items	28%
16 Items	14%
20 Items	14%

Note: While each passage had a different number of items associated with it in the item pool, six items (whether pre-determined or adaptively chosen) were always administered with each passage.

TABLE 3: Descriptive Statistics for the estimated θ Values across Ten Replications

Condition	Grand Mean	Standard Error of the Mean	Minimum Mean	Maximum Mean
Fixed	0.010	0.037	-0.038	0.074
Adaptive with MI	-0.001	0.036	-0.054	0.068
Adaptive with PR	0.000	0.037	-0.050	0.062

Note: Each replication consisted of 1,000 observations;

Known thetas: grand mean = -0.003, standard error of the mean = 0.032, minimum mean = -0.063, and maximum mean = 0.065

TABLE 4: Descriptive Statistics of the Standard Deviation of the estimated θ values across Ten Replications

Condition	Mean	Minimum	Maximum
Fixed	0.878	0.830	0.906
Adaptive with MI	0.894	0.871	0.935
Adaptive with PR	0.896	0.860	0.942

Note: Each replication consisted of 1,000 observations;
Standard deviation of Known Thetas: mean = 0.998, minimum = 0.962, and maximum = 1.037

TABLE 5: Descriptive Statistics of the Standard Errors across Ten Replications

Condition	Mean	Minimum	Maximum
Fixed	0.377	0.375	0.379
Adaptive with MI	0.310	0.309	0.312
Adaptive with PR	0.315	0.312	0.318

Note: Each replication consisted of 1,000 observations

TABLE 6: Measurement Accuracy Statistics averaged across Ten Replications

Condition	Correlation	Bias	RMSE	AAD
Fixed	0.902	-0.013	0.431	0.343
Adaptive with MI	0.920	-0.002	0.392	0.310
Adaptive with PR	0.918	-0.003	0.396	0.315

Note: Each replication consisted of 1,000 observations

TABLE 7: Descriptive Statistics of the Mean and Standard Deviation of Testlet Exposure Rates across Ten Replications

Condition	Grand Mean	Mean Min	Mean Max	Mean Std ^a	Min Std ^a	Max Std ^a
Fixed	0.040	0.000	0.301	0.056	0.056	0.057
Adaptive with MI	0.040	0.000	0.301	0.057	0.056	0.058
Adaptive with PR	0.040	0.000	0.301	0.057	0.057	0.058

Note: Each replication consisted of 1,000 observations; Std^a: Standard Deviation

TABLE 8: Pool Utilization and Frequency of Testlet Exposure Rates averaged across Ten Replications

Exposure Rate	Fixed	Adaptive with MI	Adaptive with PR
0.36 - 1.00	0	0	0
0.31 - 0.35	2	1	1
0.26 - 0.30	3	4	3
0.21 - 0.25	2	3	3
0.16 - 0.20	2	2	2
0.11 - 0.15	9	8	9
0.06 - 0.10	22	20	20
0.01 - 0.05	136	138	138
Not Administered	1	1	0
% of pool not administered	0.29	0.34	0

Note: Each replication consisted of 1,000 observations

TABLE 9: Descriptive Statistics of the Mean and Standard Deviation of Item Exposure Rates across Ten Replications

Condition	Grand Mean	Mean Min	Mean Max	Mean Std ^a	Min Std ^a	Max Std ^a
Fixed	0.017	0.000	0.301	0.042	0.042	0.042
Adaptive with MI	0.017	0.000	0.300	0.037	0.037	0.038
Adaptive with PR	0.017	0.000	0.201	0.034	0.034	0.034

Note: Each replication consisted of 1,000 observations; Std^a: Standard Deviation

TABLE 10: Pool Utilization and Frequency of Item Exposure Rates averaged across Ten Replications

Exposure Rate	Fixed	Adaptive with MI	Adaptive with PR
0.36 - 1.00	0	0	0
0.31 - 0.35	10	1	0
0.26 - 0.30	15	13	0
0.21 - 0.25	13	16	7
0.16 - 0.20	12	14	40
0.11 - 0.15	56	40	41
0.06 - 0.10	131	131	136
0.01 - 0.05	817	1422	1468
Not Administered	1367	782	727
% of pool not administered	56.49	32.33	30.04

Note: Each replication consisted of 1,000 observations

TABLE 11: Descriptive Statistics of the Overall Testlet Overlap across Ten Replications

Condition	Overall		
	Grand Mean	Minimum Mean	Maximum Mean
Fixed	0.832	0.822	0.840
Adaptive with MI	0.845	0.830	0.855
Adaptive with PR	0.846	0.833	0.861

Note: Each replication consisted of 1,000 observations

TABLE 12: Descriptive Statistics of Testlet Overlap for Examinees of Similar and Different Proficiency Levels across Ten Replications

Condition	Similar			Different		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Fixed	1.014	0.993	1.038	0.637	0.619	0.655
MI	1.054	1.024	1.077	0.618	0.598	0.641
PR	1.059	1.028	1.084	0.616	0.600	0.633

Note: Each replication consisted of 1,000 observations; “Similar” examinees were those whose known θ values differed by one logit or fewer; “Different” examinees were those whose known θ values differed by more than one logit

TABLE 13: Descriptive Statistics of the Overall Item Overlap across Ten Replications

Condition	Overall		
	Grand Mean	Minimum Mean	Maximum Mean
Fixed	4.994	4.930	5.039
Adaptive with MI	4.057	3.992	4.109
Adaptive with PR	3.465	3.427	3.524

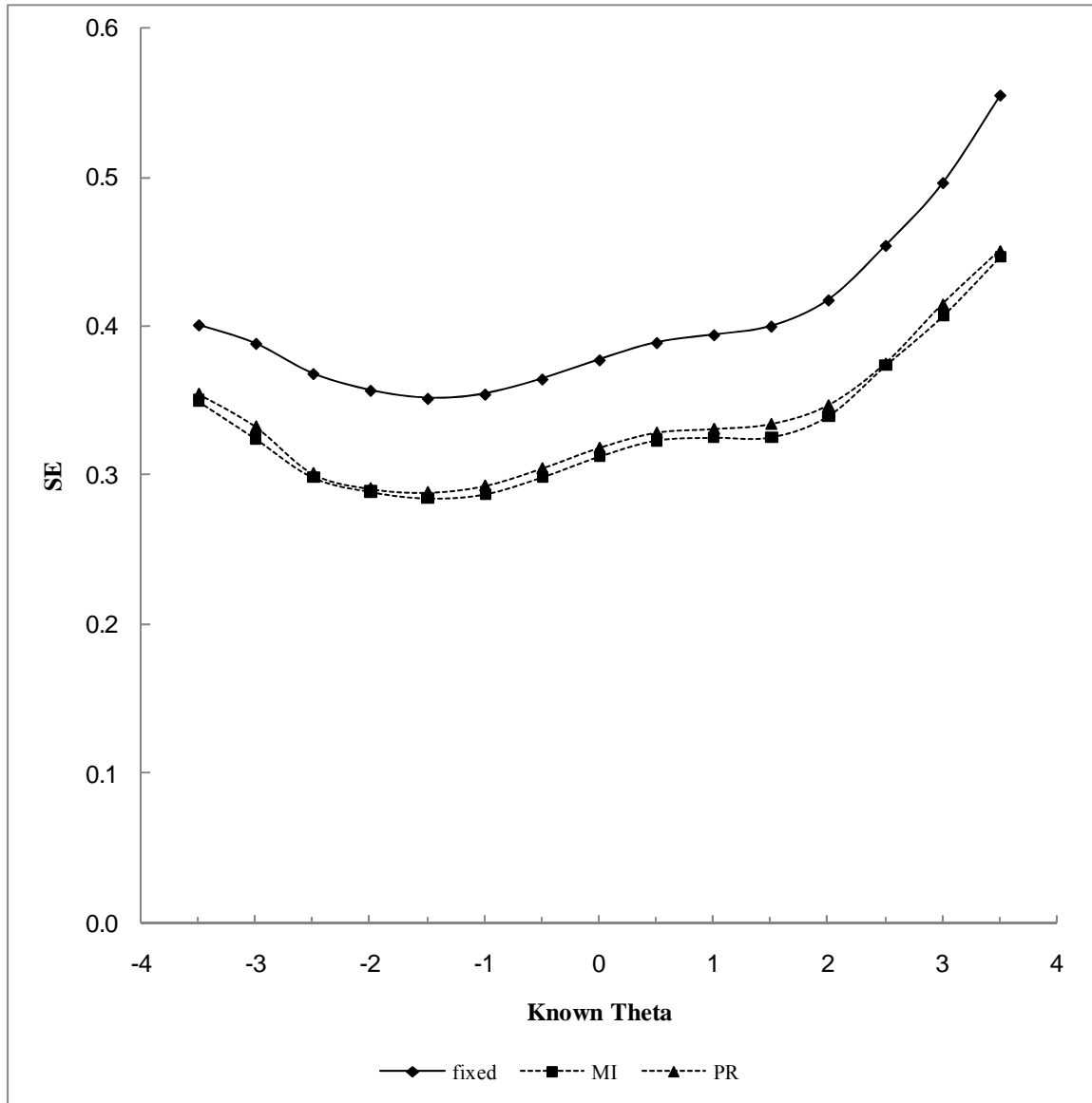
Note: Each replication consisted of 1,000 observations

TABLE 14: Descriptive Statistics of Item Overlap for Examinees of Similar and Different Proficiency Levels across Ten Replications

Condition	Similar			Different		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Fixed	6.082	5.959	6.227	3.820	3.712	3.931
MI	5.375	5.206	5.493	2.634	2.513	2.745
PR	4.541	4.450	4.661	2.305	2.241	2.375

Note: Each replication consisted of 1,000 observations; “Similar” examinees were those whose known θ values differed by one logit or fewer; “Different” examinees were those whose known θ values differed by more than one logit

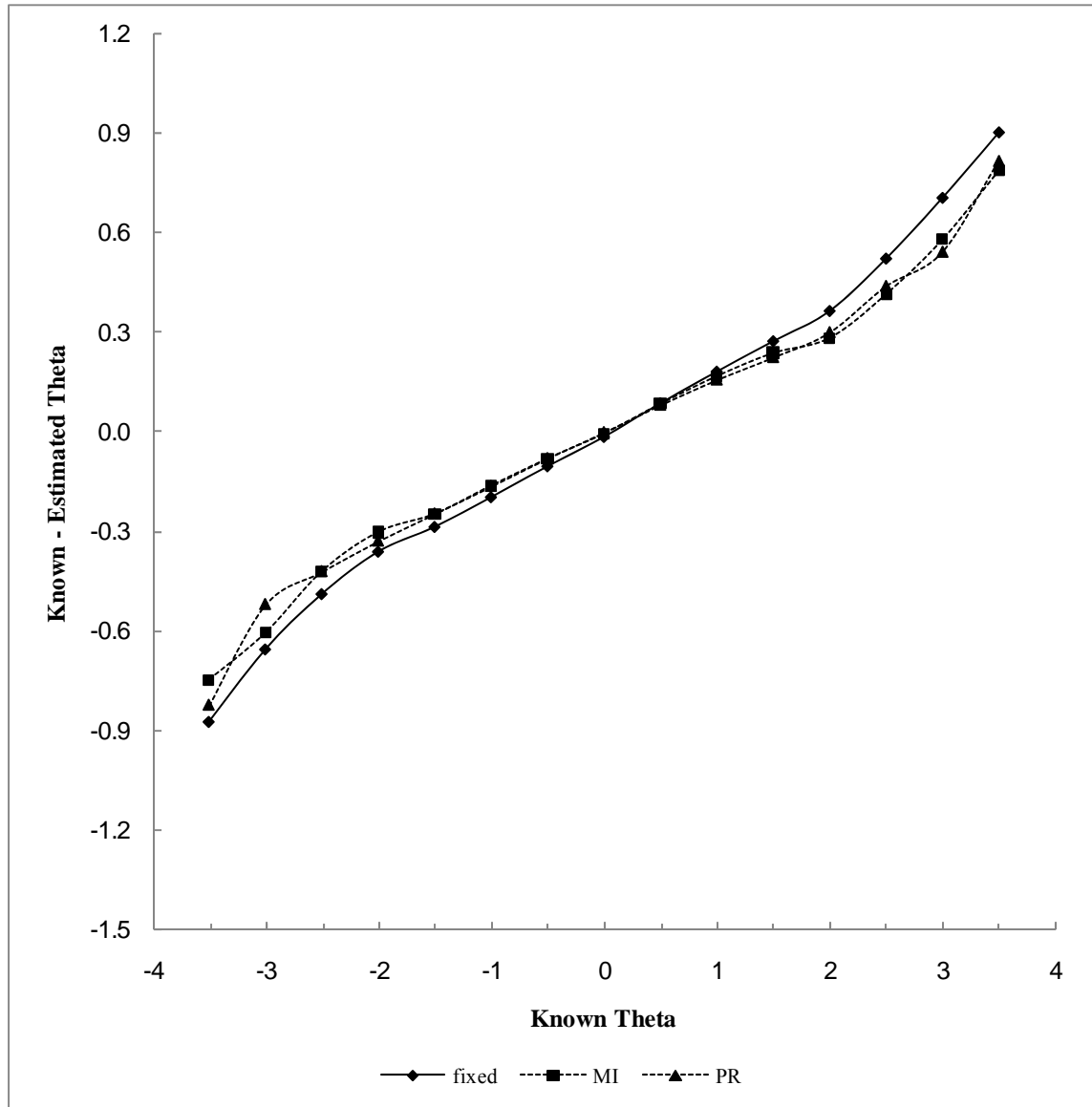
FIGURE 1: Grand Mean of the Standard Error (SE), Conditional on Known θ for the three CAT Conditions



Legend

- Fixed: CAT with fixed testlets
- MI: CAT with adaptive testlets, no item-level exposure control
- PR: CAT with adaptive testlets, with item-level exposure control

FIGURE 2: Grand Mean of the Difference between Known and Estimated θ , Conditional on Known Theta for the three CAT Condition



Legend

- Fixed: CAT with fixed testlets
- MI: CAT with adaptive testlets, no item-level exposure control
- PR: CAT with adaptive testlets, with item-level exposure control