International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 12, Number 1, February 2025
DOI 10.7333/2502-1201054
ISSN 2165-6592

The Impact of Item Bank Size and Item Bank Distribution on Student Ability Estimates for a Hybrid Interim-Summative CAT

Garron Gianopulos and Jonghwan Lee
NWEA

Sangdon Lim, Luping Niu,
Sooyong Lee, and Seung W. Choi
University of Texas at Austin

This paper investigated the impact of a uniform versus a bell-shaped distribution of items in a computerized adaptive bank within and across administrations for a hybrid interim-summative assessment. Item bank sizes of 500, 800, and 1,500 were simulated for both distributions. One-hundred simulations were conducted for two grades (Grade 4 and Grade 6) in mathematics. The item banks were generated under a Rasch model for dichotomous items and a partial credit model for three-category items. Each item bank was simulated to be vertically scaled and vertically articulated across grades. The items in the banks were generated to align with the blueprints for a state test, and the targeted distribution of items across the four performance levels was implemented based on the intended score interpretations. For the two item banks under the normal distribution, the difficulties for Grades 4 and 6 were drawn from a normal distribution with means of –0.40 and 0.40 and standard deviation of 1.1. For the two item banks under the uniform distribution, the difficulties were drawn from a uniform distribution with differing minimums and maximums for each grade: –3.6 to –2.8 for the minimums, and 2.4 to 3.2 for the maximums. The outcome variables investigated were measurement precision (i.e., root mean square error, measurement accuracy, item bank adaptivity, classification accuracy, and item exposure rate. Either item distribution generally worked well with slightly improved results for larger banks. Item bank sizes of 800 did not perform materially differently than bank sizes of 1,500. In general, while all three administrations had robust findings with the outcome variables, the

measurement quality degraded only slightly in the 500-item bank. Implications and trade-offs in item bank composition are discussed from a measurement and financial perspective.

Keywords: *classification accuracy, computerized adaptive tests, item banks, optimal item bank characteristics, through-year assessments*

The purpose of this study was to define an ideal item bank size and distributional shape for a hybrid interim-summative computerized adaptive test (CAT) that is administered three times within a school year. The purpose of this new assessment design, commonly referred to as a through-year assessment, is twofold: to provide timely and useful feedback to students and teachers throughout the school year and to provide an end-of-year summative determination for accountability. A major challenge with any CAT is the development of optimal item banks that support maximum adaptivity. CATs are maximally adaptive when items are chosen that closely correspond to the student's true ability. The more items chosen that are proximal to the student's estimated ability, the more efficient the CAT. Efficient CATs converge on the ability estimate faster with more precision than non-efficient CATs. If there are an insufficient variety and quantity of items near a student's true ability level, the CAT will be less efficient. This might also contribute to measurement error. In a CAT context, the quality and size of the item bank greatly determines the precision and accuracy of scores and resulting score inferences. Under the Rasch (1960) model, the largest factor in determining the item discrimination is the item's proximity on the scale to the classification cutscores. While a diversity of items and item difficulties is always desirable in a CAT item bank, an ideal item bank will be distributed to maximize information at the most important cutscores proportional to the student population density. Therefore, an ideal item bank will have enough items and a distributional shape that supports classification decisions for a given population distribution.

Item Bank Size

CAT experts have provided general guidance on the needed item bank size for a CAT. For example, a conventional rule of thumb for test developers who want to transition from linear test forms to CAT has been that a CAT item bank should have enough items to construct 5–10 linear test forms (Parshall et al., 2002; Stocking, 1994). Using this rule of thumb, a CAT item bank would need 200–400 items to support a single administration of a 40-item CAT (40 × 5 = 200; 40 × 10 = 400). For a CAT that is to be administered three times per year, these estimates would need to be tripled, bringing the total to 600 – 1,200 items. In light of the ongoing risk of coordinated item harvesting (see Surjadi & Randazzo, 2024; Reuters, 2016), CAT programs may mitigate such risk by sequestering one item bank from operational use that can be used to replace compromised items, similar to the use of breach forms (ITC, 2014).In this case, an additional 200 – 400 items would be necessary, bringing the total to 800 – 1,600 items.

The size of the item bank also depends on the complexity of the blueprint constraints. The larger the number of constraints, the larger the item bank that is required (Davey, 2011). Test blueprints are developed to specify the construct being measured and ensure that the construct maintains coherence and equivalence across time and across students. Blueprints vary in level of

specification. Summative tests are historically more constrained than interim or formative tests. Typically, English language arts (ELA) tends to be more constrained than mathematics because the former includes passage sets that complicate item selection. Given Davey's advice and the relatively constrained blueprints used in this simulation study, it would be expected that the size of the item bank should be closer to 1,600 items rather than 800. However, the rule of thumb approach does not provide enough certainty in exactly how many items are needed for a given CAT and set of constraints. Therefore, Davey (2011) recommends conducting simulations to reduce uncertainty.

Reducing uncertainty in the size of the required item bank is important because the costs of making an 800-item bank versus a 1,600-item bank are dramatic. According to Rudner (2009), the cost to produce a single high-quality item for a high-stakes assessment ranges from $1,500 to $2,000. If each item costs $1,500 to develop and field test, the estimated cost would be $1.2 million for an 800-item bank and $2.4 million for a 1,600-item bank. Such a large range makes it difficult to project costs with certainty. Planning for the higher cost estimate might make proposals less competitive when compared to other proposals; on the other hand, if the minimum or mean is used, costs might be underestimated and the CAT system being promised in a proposal might not be sufficiently funded. Either case is undesirable.

Item exposure rules also impact the required size of an item bank. One consideration is whether to require a one-to-one or one-to-many relationship between items and test administrations. For example, a one-to-one relationship would mean that items from the field tested and calibrated item bank would be assigned to only one season (fall, winter, or spring). The rationale for dividing the larger item bank into season-specific partitions is to prevent item exposure and reduce the risk of cheating, because if a breach occurred in the fall or winter test event, the breached items could not be used in the spring. A one-to-one relationship between item and test administration is the most conservative and most secure, but it is also the costliest approach because it would likely require a larger item bank than a one-to-many approach. A larger number of items would probably be needed to ensure that feasible solutions can be obtained in each test administration because each season-specific partition would be smaller than one large item bank.

A one-to-many relationship between item and test administration is a less conservative approach, allowing items to be used at any test administration during the school year. The main benefit to this approach is that the size of the item bank would be maximally large during each CAT administration, allowing the item selection routine to more easily find feasible solutions. The minimum item exposure control in this case would be to prevent the same student from seeing the same item during any later test event. This exposure rule would ensure that a given student would not see the same item twice within or across test events, although items might be used repeatedly within the same classroom. If a breach occurred using this approach, there is a risk that students could cheat and raise their scores artificially in the winter or spring test event.

This study was focused on the one-to-many approach and used within-person item exposure controls to prevent the same student from seeing the same item twice. While there are risks of item bank breaches (detected or not), it can be argued that the risks are mitigated by two factors. First, in the case of an undetected item breach, even if a portion of a very large item bank is compromised, the risk that a given student could benefit from knowledge of the breached items is small given the size of the item bank. For example, consider if 40 out of 800 items were breached by one cheating student and that student shared those breached items with other students. For these

students to actually benefit from these items, they would need to share a very similar ability estimate to the original student to even have a chance of seeing the items that were breached. Second, in the case of a known item bank breach, if a 40-item breach occurred, removing the 40 compromised items from the item bank is an option and should not prevent the CAT from working effectively. All of this is predicated on the notion that the item bank is large enough.

Item Bank Shape

Differently shaped item bank distributions serve different goals. For example, if the goal of a through-year CAT is to classify students into pass/fail performance categories, a high density of items is needed surrounding the cutscore (Luecht, 2006). Such an item bank would maximize information near the cutscore and reduce measurement error. A prior simulation study suggested that a test information function (TIF) of 24 near a cutscore would produce a classification accuracy rate near 0.95 (Luecht, 2006). In contrast to classifying into pass/fail categories, if the goal of a through-year CAT is to measure growth using gain scores, many items are needed all along the score continuum. The TIF for such a test would be wide but not as deep. Ideally, a uniformly shaped item bank would support a CAT optimally in its effort to produce equally precise scores all along the continuum, which, in turn, would support growth inferences. Growth inferences are known to be unreliable (Cronbach & Furby, 1970; Castellano & McCaffrey, 2020), therefore score precision is paramount if growth is the priority. Given the design of the through-year CAT, classification accuracy is paramount for both routing decisions and classifying into achievement levels, but the classification decisions include multiple cutscores spaced across the score continuum. This implies that a more uniform distribution that reaches a minimum TIF of 24 near each cutscore would be optimal for a through-year assessment. Therefore, it is necessary to conduct simulation studies to estimate needed sample sizes and examine the effect of the distributional shape of the item bank on score accuracy and item exposure. By reducing the uncertainty in the distributional shape of the item bank and the size of the item bank, test developers can project costs more accurately and devise better CAT development and maintenance plans.

Research Questions

What effect does CAT item bank size have on measurement precision, score accuracy, item bank adaptivity, classification accuracy, and item exposure of Grade 4 and Grade 6 mathematics CATs?

What effect does the distributional shape of a CAT item bank have on measurement precision, score accuracy, item bank adaptivity, classification accuracy, and item exposure of Grade 4 and Grade 6 mathematics CATs?

Method

The Modeled Item Bank

Data were simulated to mirror a pre-existing end-of-grade CAT in mathematics from Grades 3–8 used for a state accountability test. The CAT was fixed length with 41 operational items,

including seven non-operational test items that were either field test or linking items used for equating. The data were collected in Spring 2018. The item types included primarily multiple-choice items as well as some technology-enhanced items. Most items were scored dichotomously, but a small percentage were designed to be scored as polytomous items. The items were written to align to the state's content standards. An alignment study was conducted to align each item to the range achievement level descriptors within each content standard (Schneider et al., 2021).

Assumptions

No single study can include every important aspect of a complex assessment system. Therefore, it was necessary to limit the scope of this study by making certain assumptions that can be evaluated in separate studies. For example, this study could be repeated to see if results are sensitive to violations of these assumptions. The first assumption of this study was that the item bank was vertically scaled, vertically smoothed, and vertically articulated. Vertical smoothing means the item bank was trimmed, if needed, so that the minimum and maximum item difficulties were monotonically increasing across grades. This step involved removing a handful of items in the tails that contradicted the assumption of across-grade monotonicity. The reason for this trimming was to ensure that off-grade adaptivity resulted in improved score precision. Vertically articulated means that the blueprint categories in adjacent grades contained all or mostly the same domain labels. The reason for this was to help users interpret scores by combining information across grades. For example, for a student who moved off-grade after Part 1 of the test, Geometry items can be combined across grades to produce a Geometry subscore only if Geometry items are available in adjacent grades.

This study also assumed that the scale within a grade met the assumption of scale invariance and that the factor structure did not change across time. Prior studies have suggested that variations in pacing guides and instructional sequencing of content standards can cause differences in opportunity to learn (OTL; Chen, 2012). Differences in OTL might manifest in the underlying ability making it multidimensional. This type of multidimensionality can be detected using differential item functioning (DIF) where districts, schools, or teachers are treated as the grouping variable. Researchers have pointed out that if items are instructionally sensitive, they are likely to manifest DIF (Naumann et al., 2019). This time-varying OTL DIF is a threat to the IRT assumptions of unidimensionality and scale invariance.

Although time-varying OTL DIF sounds like a serious threat to the validity of this approach, prior studies have investigated the robustness of IRT scoring to the violations of these assumptions. Many studies concluded that DIF or drift had small effects on ability estimates (Wells et al., 2002; Rupp & Zumbo, 2003, 2004). One study found more meaningful effects under the two-parameter logistic model, but "for the [one-parameter logistic] 1PL model, growth parameters under the DIF and no DIF conditions were similar" (Kim & Camilli, 2014, p. 12). These studies provide evidence of the reasonableness of the assumption of unidimensionality and scale invariance. In light of these findings, this study was conducted under the assumption that scale invariance will hold.

IRT Models

At the time of simulating the item bank, the modeled item bank was in the third year of use. The Rasch model and partial credit model were originally used to calibrate the item bank and set the scale. Under the Rasch model, the probability of a student with ability θ responding correctly to item i is:

$$P\left(u_{ij}=1\middle|\theta_j,b_i\right)=\frac{\exp\left(\theta_j-b_i\right)}{1+\exp\left(\theta_j-b_i\right)} \tag{1}$$

where θ_j and b_i are the person and item parameters, respectively. Under the partial credit model, the probability of a student with ability θ having a score at the *k*th level of item i is:

$$P\left(u_{ij}=k\middle|\theta_j\right)=\frac{\exp\left[\sum_{u=1}^{k}\mathrm{D}a_i\left(\theta_j-b_i+d_{iu}\right)\right]}{\sum_{u=1}^{m_i}\exp\left[\sum_{u=1}^{k}\mathrm{D}a_i\left(\theta_j-b_i+d_{iu}\right)\right]} \tag{2}$$

where k is the score on the item (1, 2, …), m_i is the total number of score categories for the item, d_{iu} is the threshold parameter for the threshold between scores u and u–1, D=1.7, and θ_j and b_i are the person and item parameters, respectively.

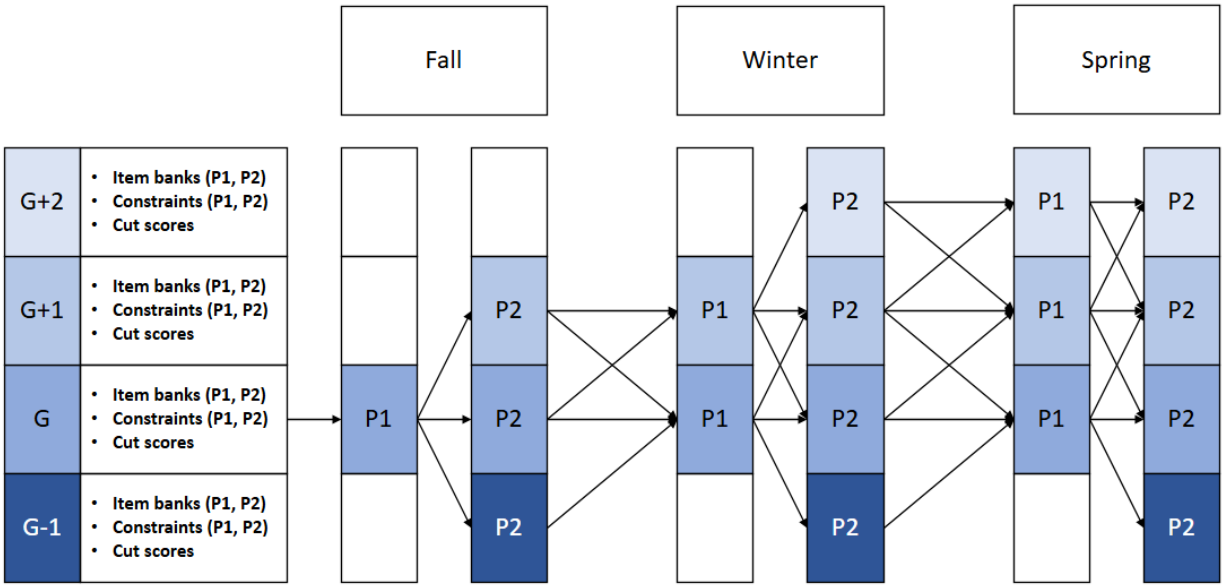
CAT Design

The CAT design was a multi-phase CAT that adapts at multiple levels, including within phases, between phases, and between tests (Figure 1). Items were adaptively selected within phases using the shadow-test approach to CAT (van der Linden & Reese, 1998). The MAAT package (Choi et al., 2022) was used to simulate three test administrations (fall, winter, spring) with two phases (P1 and P2) within each administration.

Although the flowchart in Figure 1 resembles a conventional multistage adaptive test which adapts blocks of items rather than individual items, MAAT adapts across modules and was also a fully item-level adaptive test within each module that contains a distinct item bank. The adaptive modules were configured differently across the administrations. Phase 1 of the fall administration consisted of a single module and Phase 2 consisted of three modules (i.e., a 1–3 design) reflecting different grade levels. The Phase 1 module was an adaptive routing module used to determine which module in Phase 2 was most appropriate for the student. The Phase 2 modules differed by grade, allowing students to stay on-grade, or adapt one grade below or one grade above. The winter administration had a 2–4 module design, allowing on-grade or above-grade students to continue where they left off. Students who scored below grade in the Fall, began with the on-grade module. In this way, these students were evaluated against on-grade standards. The modules in Phase 2 allowed students to adapt below or above grade. The spring administration had a 3–4 module

design, allowing students who scored above grade previously to start one or two grade levels above their grade of record. Students who scored below grade at the winter administration began the spring administration with the on-grade module, ensuring they were evaluated against on-grade standards.

Figure 1
Routing of Modules



G = Grade level, P1 = phase 1, P2 = phase 2.

Simulation Procedures

Item parameters. Two types of distributions were used to generate the item parameters: normal and uniform. Table 1 presents the mean and standard deviation for the normal distribution, and Table 2 presents the lower and upper boundary for the uniform distribution. To ensure that the vertical scale progressed across grades, the lower boundary was adjusted as needed. For example, the lowest possible b parameter for Grade 3 was –4.0, and the lowest possible b parameter for Grade 4 was –3.6. Although Grades 3 – 8 are shown in Tables 1 and 2, this simulation study was performed only on Grade 4 and Grade 6; the tables include all grades because the CAT in this study allows for off-grade items to be used. Histograms of the difficulty parameters from each distribution are shown in Appendix A. For dichotomous items, the *b* parameter was directly drawn from the distributions with specified input arguments from Table 1 and Table 2.

When modeling the polytomous items, some assumptions were made: (a) all polytomous items had three score points (0, 1, and 2), resulting in two step parameters; and (b) the following equations held:

$$b \text{ parameter} = \frac{Step\ 1 + Step\ 2}{2} \tag{3}$$

Solving Equation 3,

$$Step\ 2 = b \text{ parameter} \times 2 - Step\ 1 \quad (4)$$

Table 1
Mean and SD for Normal Distribution

Grade	Mean	SD	Lowest *b*-Parameters
3	-0.8	1.1	-4.0
4	-0.4	1.1	-3.6
5	0.0	1.1	-3.2
6	0.4	1.1	-2.8
7	0.8	1.1	-2.4
8	1.2	1.1	-2.0

Table 2
Lower and Upper Boundary for Uniform Distribution

Grade	Lower Boundary	Upper Boundary
3	-4.0	2.0
4	-3.6	2.4
5	-3.2	2.8
6	-2.8	3.2
7	-2.4	3.6
8	-2.0	4.0

Based on these assumptions, the following steps were conducted for polytomous items:

1. Randomly draw the *b* parameter from the distributions (same as for dichotomous item)s.
2. Randomly draw the Step 1 parameter from a normal distribution with mean of 0 and standard deviation of 1.
3. Calculate the Step 2 parameter using Equation 4.
4. Check for the following scenarios:
   a. If Step 1 is bigger than Step 2, discard and repeat Step 1.
   b. If the distance between Step 1 and Step 2 is larger than 1 or less than 0.2, discard and go back to Step 1.
   c. If the Step 2 parameter meets the requirements, retain it.

Ability distributions. Given the time and computing resources required to simulate three CAT administrations per simulee, it was necessary to limit the scope of this study to the ability distributions of two grade levels. Although limiting the simulations to two grade levels reduced the generalizability of this study, this compromise was necessary to stay within budget. Grades 4 and 6 were chosen to illustrate off-grade adaptivity. As shown in Figure 1, when G = 4, the item banks included Grades 3 through 6 (G – 1 to G + 2). When G = 6, the item banks included Grade 5 through 8. Thus, including Grades 4 and 6 allowed all item banks (3 through 8) to be utilized. The ability distributions were also simulated to mirror those of the real test data, albeit not from a through-year test. The mean θs at each grade level were used to set differences across grades. The grade-level spring-to-spring differences in mean θs were set to 0.40. The growth within grade was assumed to be linear from fall to spring and was set to 0.30 per season. The within-grade seasonal differences and across-grade spring differences were based on the observed differences averaged across grade levels in the real dataset (Appendix B). The ability distributions were assumed multivariate normal. The correlation of all three scales was set to r = 0.80 to mimic typical correlations of interim assessments.

Table 3
Descriptive Statistics of Simulated θs

Grade		Fall	Winter	Spring
4	Mean	-1.00	-0.70	-0.40
	SD	0.90	1.00	1.05
	Fall	1.00	0.80	0.80
	Winter	–	1.00	0.80
	Spring	–	–	1.00
6	Mean	-0.20	0.10	0.40
	SD	0.90	1.00	1.05
	Fall	1.00	0.80	0.80
	Winter	–	1.00	0.80
	Spring	–	–	1.00

Two grade levels (Grade 4 and Grade 6), two distributions (uniform and normal), and three item bank sizes (500, 800, 1,500) were crossed to produce 12 conditions total, as shown in Table 4. One hundred replications were created for each condition. In each replication, θ estimates were generated for 1,000 simulees, one score for each season (fall, winter, and spring) for a total of 3,000 scores using the parameters in Table 4. A unique random seed was set for each replication for reproducibility.

Software and CAT configurations. MAAT requires externally generated input files including simulated item banks for each grade and constraint files. Similar to multistage tests, MAAT uses cutscores to decide when a simulee is routed to a below-grade, above-grade, or on-grade item bank after each phase and test. The simulations in this study can be replicated using the generated θs (Table 3), generated item banks (Table 1, Appendix A, and Appendix B), test constraints

(Appendix C), item pool size and characteristics (Appendix D), and CAT settings described below. These CAT settings were held constant across all conditions:

Table 4
Number of Replications per Simulation Condition

Grade	Distribution	Item Bank Size		
		500	800	1500
4	Normal	100	100	100
	Uniform	100	100	100
6	Normal	100	100	100
	Uniform	100	100	100

1. The "overlap_control_policy" parameter in MAAT was set to "all", which prevents a single student from seeing the same item twice within the school year. Otherwise, no other item exposure controls were used.
2. The routing rules were defined in the MAAT function to allow off-grade routing by setting the "transition_policy" to "CI" and the "transition_CI_alpha" to 0.05. This means that simulees would not be routed off-grade unless the lower bound of the 95% confidence interval (CI) of the maximum likelihood estimate (MLE) fell above the upper cutscore, or the upper bound fell below the lower cutscore. The "combine_policy" was set to "always". This means the θ used for routing between tests was based on both phases regardless of item bank grade level.
3. The "cut_scores" in the MAAT function call were based on the actual scale so that the proportions of students falling into each achievement category would approximate the real data.
4. Table 5 presents the cutscores. All grades are shown because the lowest and highest cutscores within each grade were used for all routing rules. The default routing rules in MAAT were used. For more details on the routing structure, see the MAAT vignettes.

Table 5
Cutscores

Grade	*Level 2*	*Level 3*	*Level 4*
3	-1.47	-0.55	0.48
4	-1.07	-0.15	0.88
5	-0.67	0.25	1.28
6	-0.27	0.65	1.68
7	0.13	-1.05	2.08
8	0.53	1.45	2.48

To be clear, the item bank and test constraints within each condition were not randomly generated in each replication but remained constant across replications. Therefore, the results generalize well to the particular item banks modeled in this study and might not generalize to other item banks beyond the Grade 4 and Grade 6 item banks modeled in this study. The variation across replications was driven by the randomness of the θ sample and the interaction of the CAT item selection with the simulated students.

Outcome Measures

Results were evaluated in terms of θ recovery, item bank adaptation, and classification accuracy. θ recovery was evaluated using the following statistics: root mean square error (RMSE) and the bias of the θ estimates:.

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_m(\hat{\theta} - \theta)^2} \tag{5}$$

$$\text{Signed Bias} = \frac{1}{M}\sum_m(\hat{\theta} - \theta) \tag{6}$$

$$\text{Absolute Bias} = \frac{1}{M}\sum_m\left|\hat{\theta} - \theta\right| \tag{7}$$

$$\text{Conditional standard error of measurement (CSEM)} = \frac{1}{\sqrt{\text{I}(\theta)}} \tag{8}$$

Item bank adaptation was evaluated using a correlation index and a ratio index. The correlation index was the correlation between the average item difficulty and final θ estimate. The ratio index was computed as the standard deviation of average item difficulties over the standard deviation of final θ estimates.

$$\text{Correlation} = r\left(b_{\bullet j}, \hat{\theta}_j\right) \tag{9}$$

$$\text{Ratio} = \frac{SD(b_{\bullet j})}{SD(\hat{\theta}_j)} \tag{10}$$

Classification accuracy was calculated by counting the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classification decisions, then dividing each by the total number of decisions. A TP was identified if a simulee's estimated θ and their true θ reached or exceeded the cutscore for proficiency. A TN was identified if a simulee's estimated θ and their true θ both did not reach the cutscore for proficiency. A FP was identified if the simulee's estimated θ reached or exceeded the cutscore but their true θ did not. A FN was identified if the simulee's estimated θ did not reach the cutscore, but their true θ did reach or exceed the cutscore. These labels were summarized as percentages.

To examine the effects of item bank size and bank distribution on item exposure ,the following was done: a sample of 10 replications for each item bank was taken. Frequency counts were

obtained for each item within each replication. Then, the average frequency per item was obtained across the 10 replications. These mean frequencies were then binned into five frequency categories. The frequency counts in each category represent the number of unique items that were exposed, on average, across the 10 replications. The utilization rate was defined as the percentage of items used at least once. The primary focus was on on-grade item utilization.

Results

Research Questions

Question 1. The first questions asked was "What effect does CAT item bank size have on measurement precision, score accuracy, item bank adaptivity, classification accuracy, and item exposure of a Grade 4 and Grade 6 mathematics test?" The effects were small overall with few exceptions. The mean RMSE reduced slightly as item bank size increased, as shown in Table 6. The mean bias was not affected by item bank size. The average effect of item bank size on the correlation index was very small. The ratio index increased as the item bank size increased.

Table 6
Mean Effect of Item Bank Size
Across All Conditions and Test Events

Item Bank	RMSE	Bias	Correlation	Ratio
500	0.279	-0.0001	0.971	0.902
800	0.275	-0.0001	0.976	0.931
1,500	0.273	0.0004	0.977	0.942

The box and whisker plots in Figures 2 to 5 display the Table 6 results disaggregated by test number. The fall test event (i.e., Test 1, or T1) showed the largest RMSE across all item bank sizes and grades. In the 800 and 1,500 bank sizes, the mean RMSE was nearly equivalent in the winter (T2) and spring (T3). However, this leveling off effect across tests did not appear in the 500-item bank condition.

These effects are very small but suggest that the larger two bank sizes performed slightly better than the smallest bank in terms of the RMSE. Bias remained unchanged across test events, as shown in Figure 3. However, all remaining outcomes changed across test events: ratios increased (Figure 4) and correlations increased (Figure 5).

Typically, CATs perform worse as item exposure increases because fewer items are available to support optimal item selection. This result is counter-intuitive at first glance, but in this case the distribution of θ was not constant across tests. Figure 6 shows how the distribution of θ (Row b) changed across time to come into better alignment with the item bank information curve (Row a). The peak of the ability distribution at T3 (Row b) was in better alignment with the test information function of the item bank (Row a) when compared to T2 and T1. This was done by design to ensure that the end-of-year classification decisions were as high as possible, in keeping with conventional summative assessments. This is the likely reason the CAT performed better on most θ recovery measures at T3 and T2 than T1. The remainder of the panels in Figure 6 show results conditioned

on θ. All lines are coincident except at the tails of the distribution. The most noticeable differences appear in the CSEM) panel (Row c) where the 500-item bank is slightly elevated near the lower tail of the ability distribution in T2 and T3. The item bank size had no meaningful effect on classification accuracy, as shown in Table 7.

Table 7
Mean Classification Accuracy Across Item Bank Sizes

Item Bank	False Positives (FP)	True Positives (TP)	True Negatives (TN)	False Negatives (FN)
500	3.942	25.867	66.875	3.308
800	3.867	25.975	66.917	3.275
1,500	3.867	25.983	66.933	3.200

In terms of item exposure or item utilization, the mean percentage of on-grade items that were used at least one time ranged from 26.7% to 43.2%, as shown in Table 7. The 800- and the 500-item banks had very similar utilization rates of approximately 40%. The 1,500-item bank had the lowest utilization rate approaching 30%. Very few off-grade items were used with high frequencies. Most of the used above-grade items were used less than five times on average. The below-grade items tended to be more evenly distributed across the frequency bins compared to all other items.

Question 2. The second research question asked was "What effect does the distributional shape of a CAT item bank have on measurement precision, score accuracy, item bank adaptivity, classification accuracy, and item exposure of a Grade 4 and Grade 6 mathematics test?" When comparing grand means, the distributional shape of the item banks did not have an effect on the RMSE or bias of the θ estimates, as shown in Table 8. However, under closer inspection, the uniform distribution did generate slightly larger RMSEs at Grade 4, especially with the smallest item bank (Figure 2).

Figure 2
RMSE Across Tests and Conditions

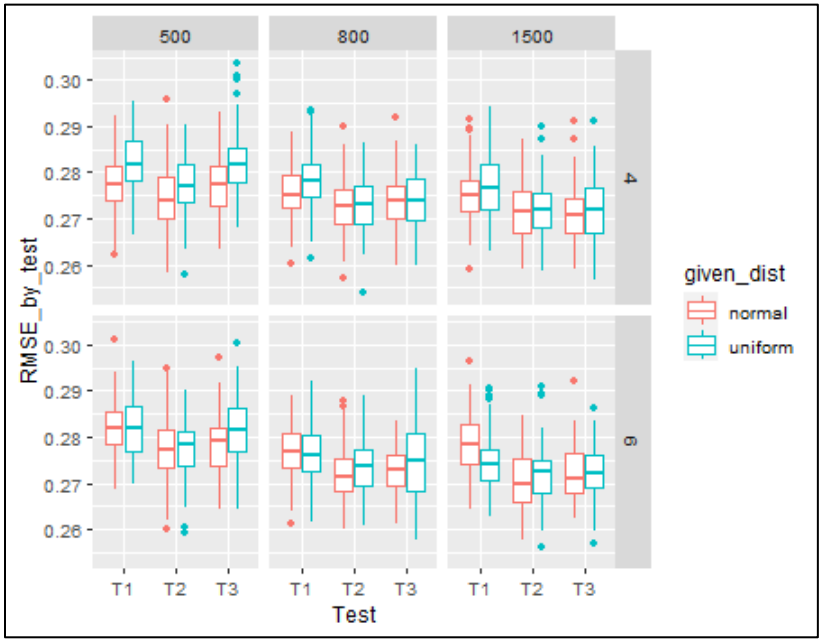


Figure 3
Bias Across Tests and Conditions

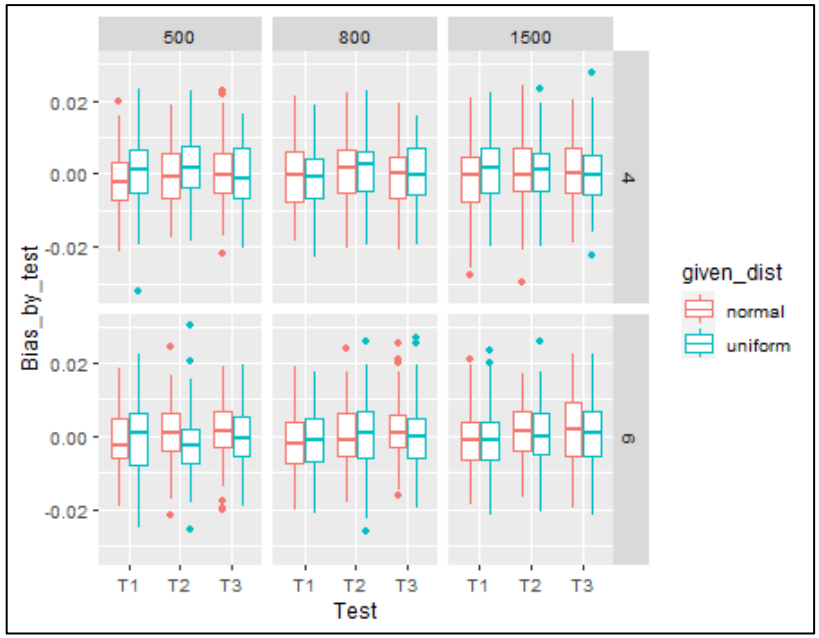


Figure 4
Ratio Index Across Tests and Conditions

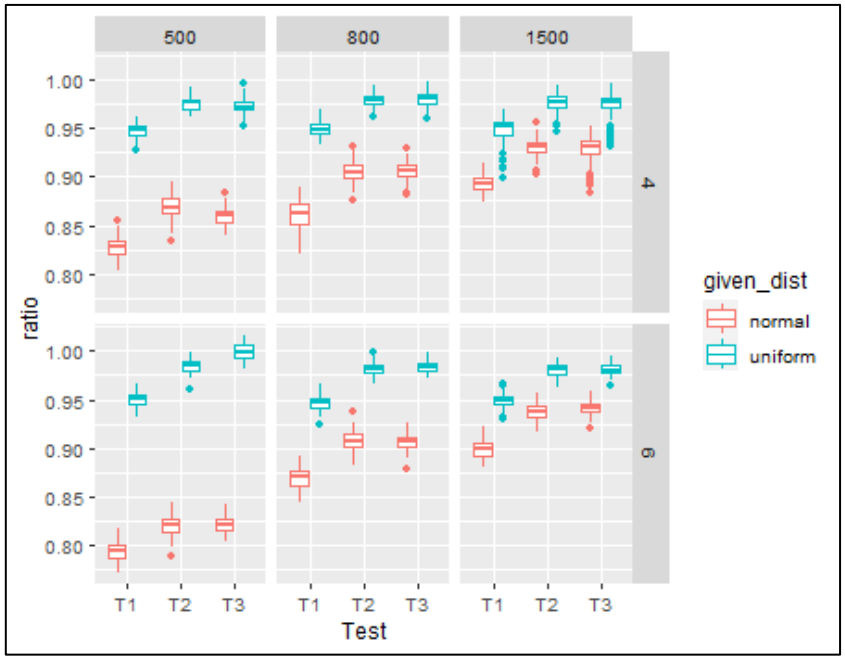


Figure 5
Correlation Index Across Tests and Conditions

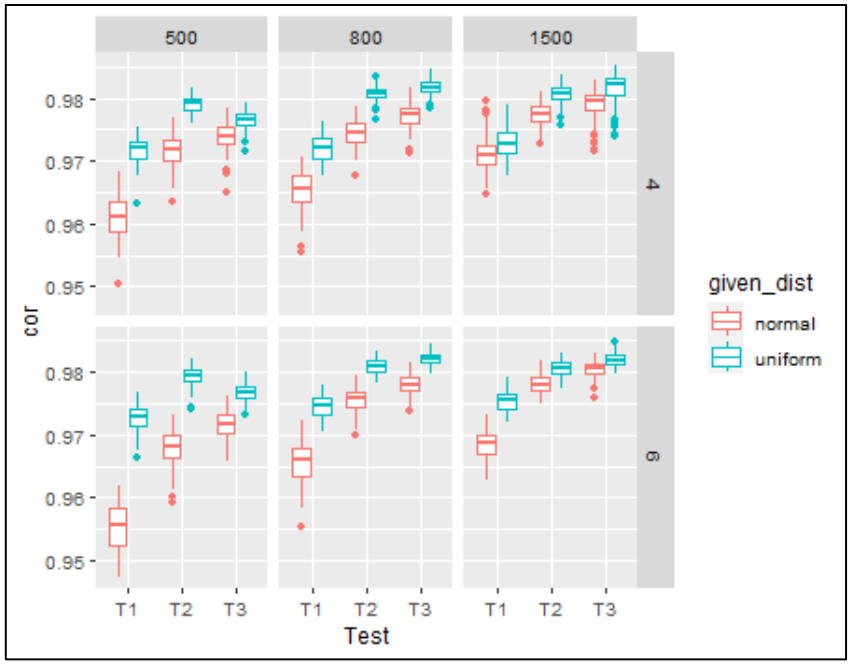


Figure 6
Effects of Item Bank Size on θ Recovery (Mean Across Conditions)

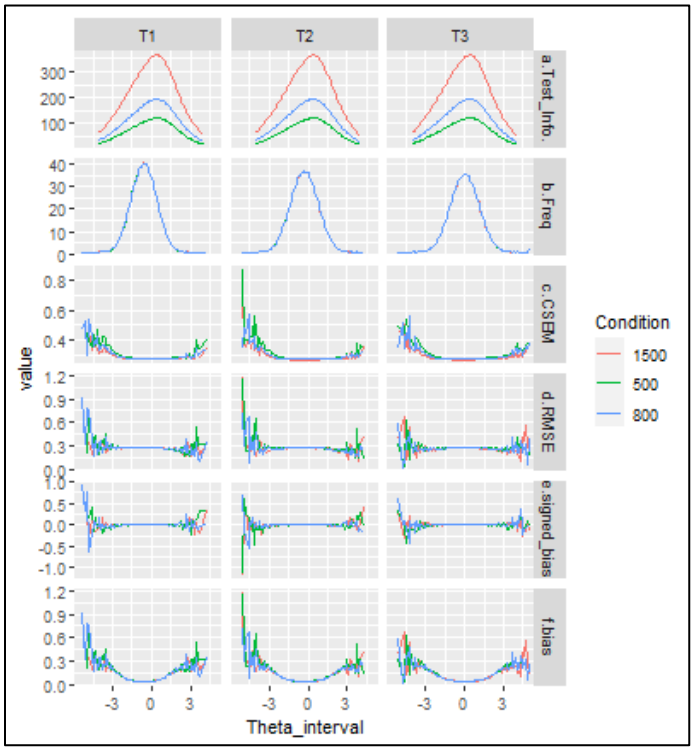


Figure 7
Effects of Item Bank Shape on θ Recovery (Mean Across Conditions)

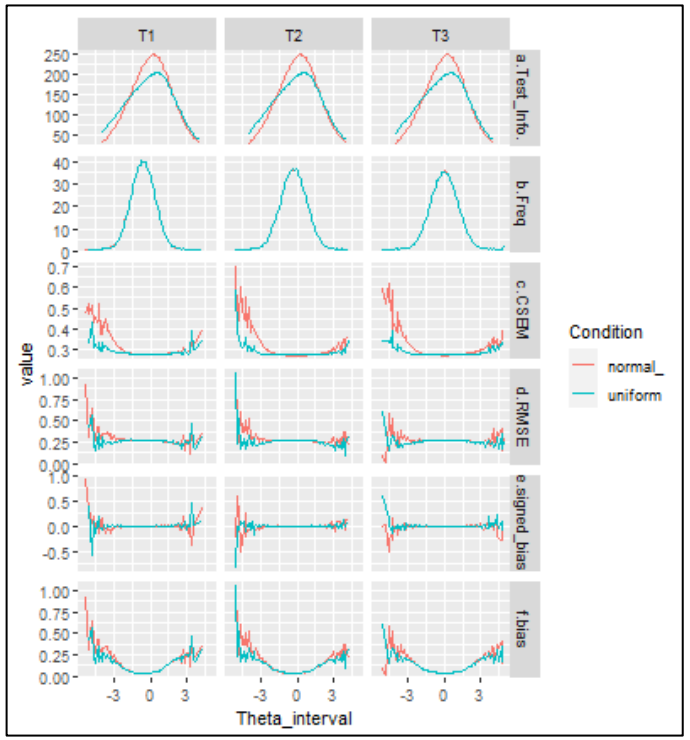


Table 8
Number of Unique Item Exposures by Frequency Category

Item Bank	Distribution	Item Bank Grade	Student Grade	Frequency Bins					Total	Utilization Rate
				1–5	6–25	26–50	51–100	>100		
500	Normal	3	4	20	6	14	18	24	82	16.4%
500	Uniform	3	4	24	22	12	24	18	100	20.0%
500	Normal	4	4	12	12	16	16	160	216	43.2%
500	Uniform	4	4	10	22	6	26	150	214	42.8%
500	Normal	5	4	96	56	0	0	0	152	30.4%
500	Uniform	5	4	92	58	2	0	0	152	30.4%
500	Normal	6	4	106	2	0	0	0	108	21.6%
500	Uniform	6	4	130	0	0	0	0	130	26.0%
500	Normal	5	6	12	10	2	30	20	74	14.8%
500	Uniform	5	6	22	22	10	28	16	98	19.6%
500	Normal	6	6	4	14	2	10	172	202	40.4%
500	Uniform	6	6	8	16	20	24	134	202	40.4%
500	Normal	7	6	78	48	0	0	0	126	25.2%
500	Uniform	7	6	80	50	2	0	0	132	26.4%
500	Normal	8	6	106	6	0	0	0	112	22.4%
500	Uniform	8	6	124	0	0	0	0	124	24.8%
800	Normal	3	4	20	4	14	28	18	84	10.5%
800	Uniform	3	4	72	24	22	30	10	158	19.8%
800	Normal	4	4	38	16	30	28	186	298	37.3%
800	Uniform	4	4	54	34	22	38	178	326	40.8%
800	Normal	5	4	112	48	4	0	0	164	20.5%
800	Uniform	5	4	126	52	0	0	0	178	22.3%
800	Normal	6	4	140	4	0	0	0	144	18.0%
800	Uniform	6	4	154	4	0	0	0	158	19.8%
800	Normal	5	6	22	12	16	36	12	98	12.3%
800	Uniform	5	6	44	24	24	38	6	136	17.0%
800	Normal	6	6	34	34	12	40	180	300	37.5%
800	Uniform	6	6	48	44	24	40	166	322	40.3%
800	Normal	7	6	120	46	2	0	0	168	21.0%
800	Uniform	7	6	116	58	0	0	0	174	21.8%
800	Normal	8	6	150	2	0	0	0	152	19.0%
800	Uniform	8	6	148	2	0	0	0	150	18.8%
1500	Normal	3	4	36	22	28	24	12	122	8.1%

Item Bank	Distribution	Item Bank Grade	Student Grade	Frequency Bins					Total	Utilization Rate
				1–5	6–25	26–50	51–100	>100		
1500	Uniform	3	4	88	32	26	28	8	182	12.1%
1500	Normal	4	4	88	58	22	36	196	400	26.7%
1500	Uniform	4	4	104	60	40	56	180	440	29.3%
1500	Normal	5	4	152	54	0	0	0	206	13.7%
1500	Uniform	5	4	174	60	0	0	0	234	15.6%
1500	Normal	6	4	188	2	0	0	0	190	12.7%
1500	Uniform	6	4	194	0	0	0	0	194	12.9%
1500	Normal	5	6	28	30	24	42	4	128	8.5%
1500	Uniform	5	6	60	38	18	36	8	160	10.7%
1500	Normal	6	6	76	50	30	56	190	402	26.8%
1500	Uniform	6	6	108	90	38	42	170	448	29.9%
1500	Normal	7	6	150	60	4	0	0	214	14.3%
1500	Uniform	7	6	148	60	0	0	0	208	13.9%
1500	Normal	8	6	164	0	0	0	0	164	10.9%
1500	Uniform	8	6	166	0	0	0	0	166	11.1%

Note. This table was made by taking a sample of 10 replications, counting the frequency each item was used, averaging the frequency counts across replications, and then binning them into five frequency bins. Utilization rate = % items used at least once.

Although the effect was small, the uniform distribution did increase the item bank adaptivity (ratio index) when compared to the normally distributed item bank from Table 9. The ratio index and the correlation index were slightly larger for the uniform distribution than the normal distribution across all conditions (Figures 4 and 5). The item distributional shape had no meaningful effect on classification accuracy, as shown in Table 10. This is evidence that TIFs in both the normal and uniform distributions were sufficient to support quality classification decisions at the proficiency cutscore.

Table 9
Mean Effects of Item Bank Distribution on Outcomes

Distribution	RMSE	Bias	Correlation	Ratio
Normal	0.275	-0.00002	0.972	0.882
Uniform	0.276	-0.00009	0.978	0.968

Table 10
Mean Classification Accuracy Across Item Bank Distributions

Distribution	False Positives (FP)	True Positives (TP)	True Negatives (TN)	False Negatives (FN)
Normal	3.872	25.967	66.917	3.250
Uniform	3.911	25.917	66.900	3.272

When comparing grand means, this study suggests that there is less benefit to using a uniform distribution than originally expected. However, if results are disaggregated across the θ distribution some effects become visible. A noticeable effect is visible at the tails of the θ distribution at each test event, especially at T2, as shown in Figure 7. The uniform distribution produced smaller CSEMs, lower RMSEs, and slightly less bias at the bottom tail and to some extent at the top tail of the θ distributions. This is due to the slightly greater amount of information in the tails of the uniform item bank.

Summary

The effects of bank size and distributional shape on θ recovery, item bank adaptation, and classification accuracy were small to negligible. Classification accuracy of the cutscores studied in this simulation was not affected by any of the studied conditions. Some consistent patterns did emerge across outcomes that are noteworthy; the two larger item bank sizes of 800 and 1,500 items tended to behave more similarly, both showing similar levels of improvement compared to the smallest item bank size of 500. Item utilization rates were very similar across conditions, with the uniform distributions showing small but consistent increases in utilization rates. The adaptivity of the item bank improved under the uniform distribution, but this beneficial effect was also small. The benefits of the uniform distribution were most visible in the tails of the θ distribution. The combination of conditions that produced the most visible negative effects in the lower tail of the θ distribution in terms of CSEM and RMSE was the normally distributed bank of 500 items, as shown in Figure 8. In contrast, the uniform distribution with 800 items or 1,500 items performed the best in terms of these metrics.

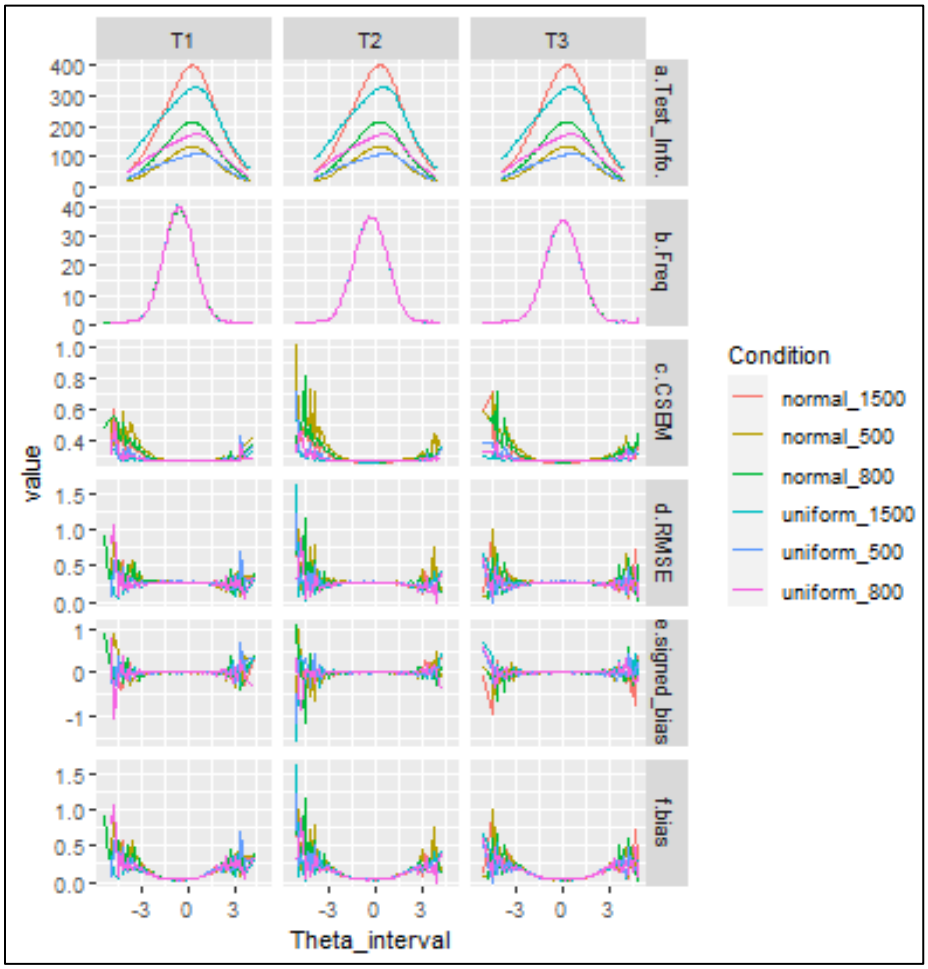
The effects of item bank distribution on on-grade item exposure were mostly inconsequential. The differences in utilization rates between the normal and uniform distributions were negligible for bank sizes of 500, as shown in Table 8. The uniform banks of the 800 and 1,500 items showed slightly higher utilization rates. The uniform distribution used a larger number of below-grade items than the normal distribution, especially at Grade 8, as shown in Table 8.

Discussion

The effects of item bank size and distribution were small to negligible. Classification accuracy was not influenced by item bank size or distribution. The item bank of 800 items only performed slightly better in RMSE than the 500-item bank. The improvements were most visible at the tails of the θ distributions, especially at the low end. The 1,500 item bank showed little to no improvement over and above the 800-item bank across any conditions. If the improvements in θ recovery were the only benefits to gain from a larger item bank, it would hardly justify the added cost of

building a bank of 800 or 1,500 items. These results demonstrate the robustness of CAT to a smaller item bank of 500 items across both item bank distributions. A benefit of using a smaller item bank of 500 items is that it would reduce costs since it reduces sample sizes required for initial item parameter calibrations and ongoing drift studies. If a testing program is required to release a substantial percentage of items annually and sample sizes and costs are less of a concern, than an item bank of 800 might be necessary and will provide a small but consistent improvement in RMSE, CSEM, and bank adaptivity.

Figure 8
Combined Effects of Item Bank Distribution and Size on θ Recovery (Mean Across Conditions)



If off-grade adaptivity was not allowed, the benefits of a uniform distribution might be greater than what was discovered in this study. Given the very small benefits of a uniform distribution of item difficulties, if off-grade adaptivity is allowed, building a uniform distribution appears to be less of a requirement. It is likely that the off-grade adaptivity provides more information than an on-grade item bank by virtue of the fact that during Part 2 of the CAT, the off-grade items can further improve the precision of the θ estimates. In contrast, an on-grade-only item bank cannot

compensate for lack of information in the tails. In this case, on-grade-only CATs would likely benefit more from a uniform item bank.

Limitations

These results generalize to the subject of mathematics with one state's item bank and blueprint constraints. The particular set of CAT constraints used in this study were moderately complex, for example, there were 48 rows of constraints in each phase of the CAT at Grade 4, as shown in Appendix C. These constraints were not complicated by the requirement to avoid long lists of item enemies or the use of common stimuli. Separate simulation studies should be completed for any tests that include more complex constraints. For example, if blueprints require item stimuli or passage sets, CAT constraints become more complex, and the risk of infeasible solutions increases. Therefore, these results do not generalize well to ELA tests that include item sets that must appear within specific passages. Furthermore, if a one-to-one relationship of item to seasonal test is desired, these recommendations would not be appropriate. In this case, a larger item bank is probably needed. Finally, this simulation study assumed an off-grade through-year CAT. Thus, if estimated item bank sizes are needed for an on-grade through-year CAT, a larger item bank might be needed so additional simulations are warranted. For these findings to transfer to Grades 3 and 8, item banks for Grade 2 and Grade 9 would need to be developed. Future studies should include simulations using ELA blueprint constraints and item exposure rules to avoid the overuse of certain items.

This present study generalizes to cases in which on-grade item difficulty is well matched to on-grade student ability, however, it is possible that operational items might be systematically too difficult for students in fall and winter for lack of OTL. Future studies should include conditions that mirror realistic levels of mismatch between item difficulty and student ability. Finally, future studies could check the sensitivity of these results to violations of the assumption of unidimensionality caused by time-varying DIF, resulting from time-varying OTL, an issue that might be unique to through-year designs.

References

Castellano, K. E., & McCaffrey, D. F. (2020). Comparing the accuracy of student growth measures. *Journal of Educational Measurement*, *57*(1), 71-91.

Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods.* Doctoral dissertation, University of Kansas. *WebLink*

Choi, S., Lim, S., Niu, L., & Lee, S. (2022). MAAT: Multiple administrations adaptive testing. R package Version 1.1.0. *WebLink*

Cronbach, L. J., & Furby, L. (1970). How we should measure "change". Or should we? *Psychological bulletin*, *74*(1), 68. *CrossRef*

Davey, T. (2011). A guide to computer adaptive testing systems. Council of Chief State School Officers. *WebLink*

Hassan, M. U., & Miller, F. (2019). Optimal item calibration for computerized achievement tests. *Psychometrika*, 84(4), 1101-1128. CrossRef

International Test Commission. (2014). International Guidelines on the Security of Tests, Examinations, and Other Assessments. WebLink

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-scale Assessments in Education, 2*(1). CrossRef

Luecht, R. M. (2006). Designing tests for pass–fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575–596). Lawrence Erlbaum Associates. WebLink

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction, 60*, 41–53. CrossRef

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing.* Springer Science & Business Media.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* MESA Press.

Reuters. (2016, June 11). Widespread cheating detailed in a program owned by test giant ACT. Reuters. WebLink

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 151–165). Springer. WebLink

Rupp, A.A., & Zumbo, B.D. (2003, April). Bias coefficients for lack of invariance in unidimensional IRT models. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588–599. CrossRef

Schneider, M. C., Agrimson, J., & Veazey, M. (2021). The relationship between item developer alignment of items to range achievement-level descriptors and item difficulty: Implications for validating intended score interpretations. *Educational Measurement: Issues and Practice*, *41*(2), 12–24. CrossRef

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item banks.* ETS Research Report RR-94-05. ETS Research Report Series, 1994(1), i–34. WebLink

Surjadi, Milla & Randazzo, Sara. (2024, August 15). The cheating scandal rocking the world of elite high-school math. *The Wall Street Journal*. WebLink

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*(3), 259–270 . CrossRef

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77–87. CrossRef

Acknowledgments

NWEA supported the development of the MAAT package and much of this research.

Author's Address

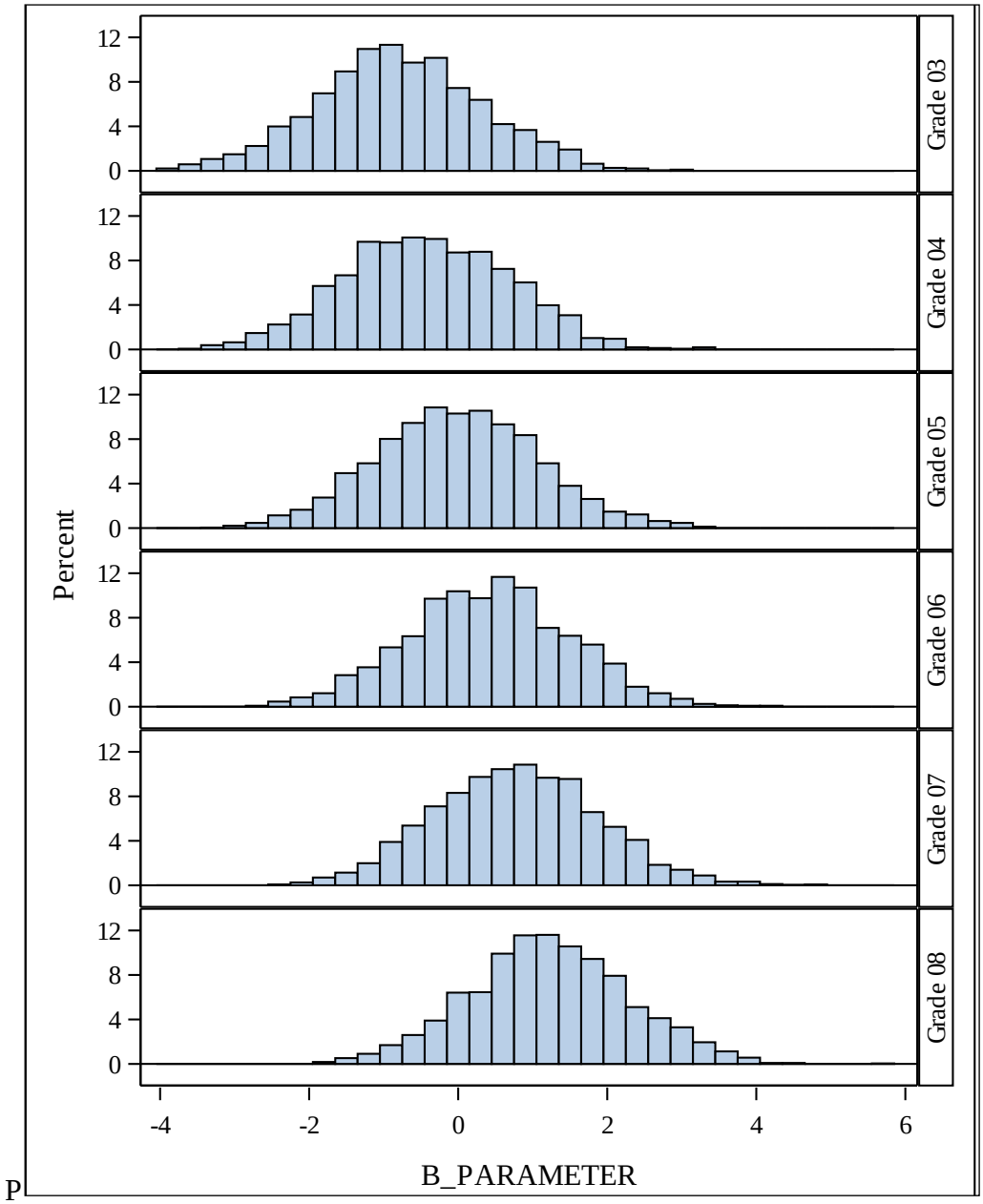
garron@gianopulos.com

Citation

G. Gianopulos, J. Lee, S. Lim, L. Niu, S. Lee, & S. W. Choi. (2025). The impact of item bank size and item bank distribution on student ability estimates for a hybrid interim-summative CAT. *Journal of Computerized Adaptive Testing, 12(1), 54-87*

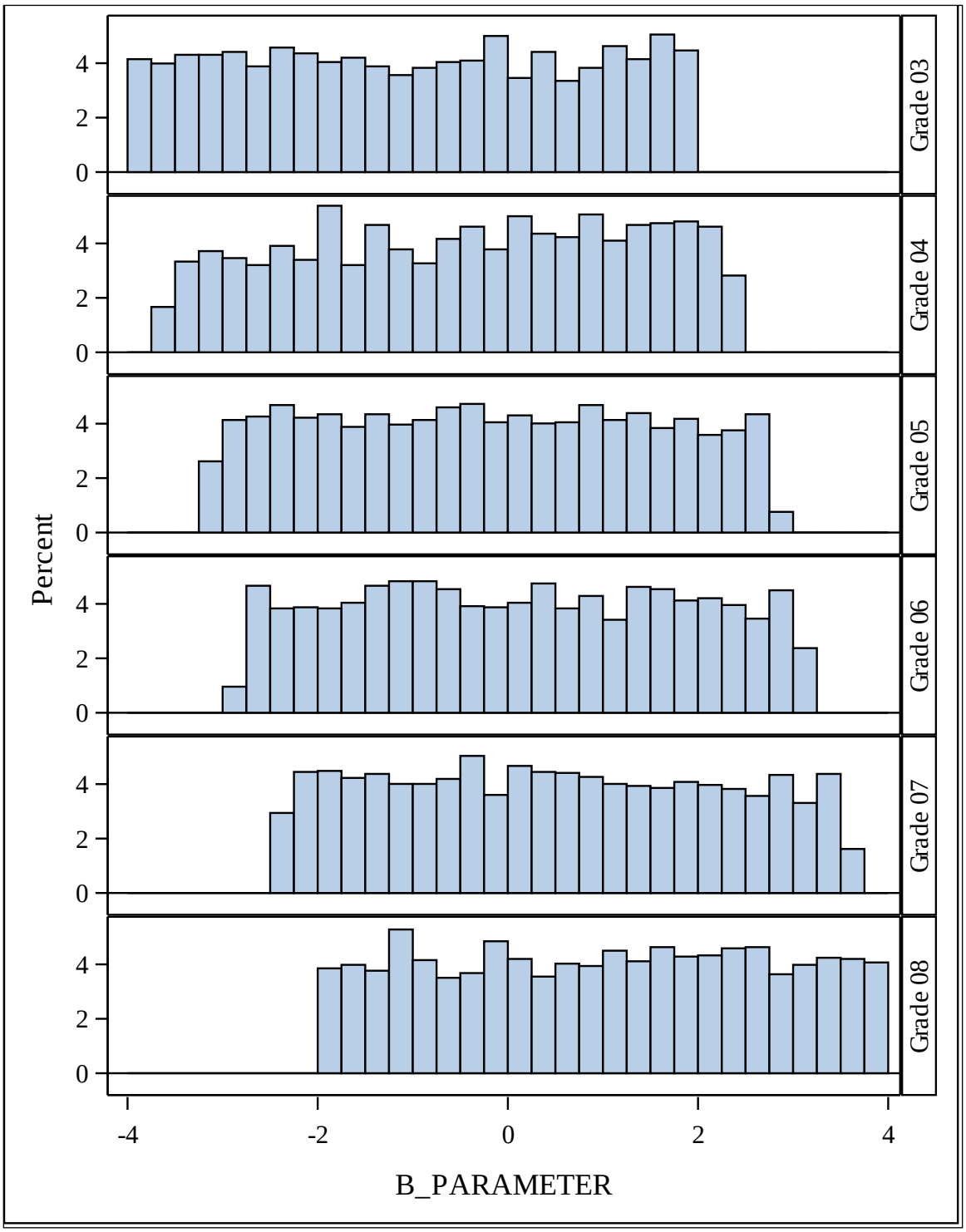
Appendix A: Simulated Item Bank Characteristics

Figure A.1
***b* Parameter Histogram Drawn from a Normal Distribution—Mathematics**



P

Figure A.2
***b* Parameter Histogram Drawn from a Uniform Distribution—Mathematics**



Appendix B: Descriptive Statistics of Vertical Scale

Table B.1
θ Means, Standard Deviations and Differences Used for Data Generation

	Mean θs			θ Standard Deviations			Differences of Means		
Grade	Fall	Winter	Spring	Fall	Winter	Spring	W-F	S-W	S-S
3	-1.33	-0.84	-0.54	0.84	0.85	0.88	0.48	0.30	
4	-0.64	-0.23	0.05	0.90	0.93	0.97	0.41	0.28	0.59
5	-0.04	0.31	0.56	0.95	0.99	1.04	0.35	0.25	0.52
6	0.31	0.61	0.82	1.01	1.05	1.09	0.30	0.21	0.26
7	0.65	0.89	1.06	1.09	1.12	1.16	0.24	0.17	0.24
8	0.95	1.15	1.29	1.18	1.21	1.25	0.20	0.14	0.22
Means	-0.01	0.32	0.54	1.00	1.03	1.07	0.33	0.22	0.37
Values used for differences in simulated θ :							**0.30**	**0.30**	**0.40**

Note. F = Fall, W= Winter, S= Spring

Appendix C: MAAT Constraint Files

Table C.1
Phase 1 Grade 4 Mathematics Constraints

Constraint ID	Type	What	Condition	LB	UB
1	Number	Item		25	25
2	Number	Item	ITEM_TYPE == "Polytomous"	4	7
3	Number	Item	STANDARD == "MA 4.1.1.a" & DOK %in% c(1, 2)	0	1
4	Number	Item	STANDARD == "MA 4.1.1.c" & DOK %in% c(1, 1)	0	1
5	Number	Item	STANDARD == "MA 4.1.1.d" & DOK %in% c(1, 1)	0	1
6	Number	Item	STANDARD == "MA 4.1.1.e" & DOK %in% c(1, 1)	0	1
7	Number	Item	STANDARD == "MA 4.1.1.f" & DOK %in% c(1, 1)	0	2
8	Number	Item	STANDARD == "MA 4.1.1.g" & DOK %in% c(1, 1)	0	1
9	Number	Item	STANDARD == "MA 4.1.1.h" & DOK %in% c(1, 1)	1	2
10	Number	Item	STANDARD == "MA 4.1.1.k" & DOK %in% c(1, 2)	0	1
11	Number	Item	DOMAIN == "NR"	4	6
12	Number	Item	STANDARD == "MA 4.1.2.b" & DOK %in% c(1, 1)	0	1
13	Number	Item	STANDARD == "MA 4.1.2.c" & DOK %in% c(1, 1)	0	2
14	Number	Item	STANDARD == "MA 4.1.2.d" & DOK %in% c(1, 1)	0	1

Constraint ID	Type	What	Condition	LB	UB
15	Number	Item	STANDARD == "MA 4.1.2.f" & DOK %in% c(1, 1)	0	1
16	Number	Item	STANDARD == "MA 4.1.2.g" & DOK %in% c(1, 1)	0	1
17	Number	Item	DOMAIN == "NO"	5	6
18	Number	Item	UDOMAIN == "NUM"	10	10
19	Number	Item	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1	10
20	Number	Item	STANDARD == "MA 4.2.1.a" & DOK %in% c(1, 2)	1	2
21	Number	Item	DOMAIN == "AR"	1	1
22	Number	Item	STANDARD == "MA 4.2.2.a" & DOK %in% c(1, 2)	2	3
23	Number	Item	DOMAIN == "AP"	2	2
24	Number	Item	STANDARD == "MA 4.2.3.a" & DOK %in% c(2, 2)	1	2
25	Number	Item	STANDARD == "MA 4.2.3.b" & DOK %in% c(2, 2)	1	2
26	Number	Item	DOMAIN == "AA"	3	3
27	Number	Item	UDOMAIN == "ALG"	6	6
28	Number	Item	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1	6
29	Number	Item	STANDARD == "MA 4.3.1.b" & DOK %in% c(1, 2)	0	1
30	Number	Item	STANDARD == "MA 4.3.1.c" & DOK %in% c(1, 2)	0	1
31	Number	Item	STANDARD == "MA 4.3.1.d" & DOK %in% c(2, 3)	0	1
32	Number	Item	STANDARD == "MA 4.3.1.e" & DOK %in% c(1, 2)	0	1
33	Number	Item	STANDARD == "MA 4.3.1.f" & DOK %in% c(1, 2)	0	1
34	Number	Item	STANDARD == "MA 4.3.1.g" & DOK %in% c(1, 2)	0	1
35	Number	Item	STANDARD == "MA 4.3.1.h" & DOK %in% c(1, 2)	0	1
36	Number	Item	DOMAIN == "GC"	4	4
37	Number	Item	STANDARD == "MA 4.3.3.a" & DOK %in% c(1, 2)	0	1
38	Number	Item	STANDARD == "MA 4.3.3.c" & DOK %in% c(1, 1)	0	1
39	Number	Item	DOMAIN == "GM"	1	1
40	Number	Item	UDOMAIN == "GEO"	5	5
41	Number	Item	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1	5
42	Number	Item	STANDARD == "MA 4.4.1.a" & DOK %in% c(2, 2)	0	2
43	Number	Item	DOMAIN == "DR"	1	2
44	Number	Item	STANDARD == "MA 4.4.2.a" & DOK %in% c(2, 2)	2	3
45	Number	Item	DOMAIN == "DA"	2	2
46	Number	Item	UDOMAIN == "DTA"	4	4
47	Number	Item	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1	4
48	SUM	Item	POINTS	29	32

Note. LB = lower bound, UB = upper bound.

Table C.2
Phase 2 Grade 4 Mathematics Constraints

Constraint ID	Type	What	Condition	LB	UB
1	Number	Item		16	16
2	Number	Item	ITEM_TYPE == "Polytomous"	4	5
3	Number	Item	STANDARD == "MA 4.1.1.a" & DOK %in% c(1, 2)	0	1
4	Number	Item	STANDARD == "MA 4.1.1.c" & DOK %in% c(1, 1)	0	1
5	Number	Item	STANDARD == "MA 4.1.1.d" & DOK %in% c(1, 1)	0	1
6	Number	Item	STANDARD == "MA 4.1.1.e" & DOK %in% c(1, 1)	0	1
7	Number	Item	STANDARD == "MA 4.1.1.f" & DOK %in% c(1, 1)	0	1
8	Number	Item	STANDARD == "MA 4.1.1.g" & DOK %in% c(1, 1)	0	1
9	Number	Item	STANDARD == "MA 4.1.1.h" & DOK %in% c(1, 1)	1	1
10	Number	Item	STANDARD == "MA 4.1.1.k" & DOK %in% c(1, 2)	0	1
11	Number	Item	DOMAIN == "NR"	3	4
12	Number	Item	STANDARD == "MA 4.1.2.b" & DOK %in% c(1, 1)	0	1
13	Number	Item	STANDARD == "MA 4.1.2.c" & DOK %in% c(1, 1)	0	1
14	Number	Item	STANDARD == "MA 4.1.2.d" & DOK %in% c(1, 1)	0	1
15	Number	Item	STANDARD == "MA 4.1.2.f" & DOK %in% c(1, 1)	0	1
16	Number	Item	STANDARD == "MA 4.1.2.g" & DOK %in% c(1, 1)	0	1
17	Number	Item	DOMAIN == "NO"	3	4
18	Number	Item	UDOMAIN == "NUM"	7	7
19	Number	Item	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1	7
20	Number	Item	STANDARD == "MA 4.2.1.a" & DOK %in% c(1, 2)	0	1
21	Number	Item	DOMAIN == "AR"	0	1
22	Number	Item	STANDARD == "MA 4.2.2.a" & DOK %in% c(1, 2)	1	2
23	Number	Item	DOMAIN == "AP"	1	2
24	Number	Item	STANDARD == "MA 4.2.3.a" & DOK %in% c(2, 2)	0	1
25	Number	Item	STANDARD == "MA 4.2.3.b" & DOK %in% c(2, 2)	0	2
26	Number	Item	DOMAIN == "AA"	1	2
27	Number	Item	UDOMAIN == "ALG"	4	4
28	Number	Item	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1	4
29	Number	Item	STANDARD == "MA 4.3.1.b" & DOK %in% c(1, 2)	0	1
30	Number	Item	STANDARD == "MA 4.3.1.c" & DOK %in% c(1, 2)	0	1
31	Number	Item	STANDARD == "MA 4.3.1.d" & DOK %in% c(2, 3)	0	1
32	Number	Item	STANDARD == "MA 4.3.1.e" & DOK %in% c(1, 2)	0	1
33	Number	Item	STANDARD == "MA 4.3.1.f" & DOK %in% c(1, 2)	0	1
34	Number	Item	STANDARD == "MA 4.3.1.g" & DOK %in% c(1, 2)	0	1
35	Number	Item	STANDARD == "MA 4.3.1.h" & DOK %in% c(1, 2)	0	1
36	Number	Item	DOMAIN == "GC"	2	2
37	Number	Item	STANDARD == "MA 4.3.3.a" & DOK %in% c(1, 2)	0	1

Constraint ID	Type	What	Condition	LB	UB
38	Number	Item	STANDARD == "MA 4.3.3.c" & DOK %in% c(1, 1)	0	1
39	Number	Item	DOMAIN == "GM"	1	1
40	Number	Item	UDOMAIN == "GEO"	3	3
41	Number	Item	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1	3
42	Number	Item	STANDARD == "MA 4.4.1.a" & DOK %in% c(2, 2)	0	2
43	Number	Item	DOMAIN == "DR"	1	2
44	Number	Item	STANDARD == "MA 4.4.2.a" & DOK %in% c(2, 2)	0	2
45	Number	Item	DOMAIN == "DA"	1	2
46	Number	Item	UDOMAIN == "DTA"	2	2
47	Number	Item	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1	2
48	SUM	Item	POINTS	20	21

Note. LB = lower bound, UB = upper bound

Appendix D: Frequency of On-Grade Items by Constraints

Table D.1
Summary of Grade 4 Item Bank Counts by Constraint (Parts 1 and 2 Combined)

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1560	834	520
1		41	41	1,560	834	520
2	ITEM_TYPE == "Polytomous"	8	12	314	172	108
3	STANDARD == "MA 4.1.1.a" & DOK %in% c(1, 2)	0	2	45	24	15
4	STANDARD == "MA 4.1.1.c" & DOK %in% c(1, 1)	0	2	45	24	15
5	STANDARD == "MA 4.1.1.d" & DOK %in% c(1, 1)	0	2	45	24	15
6	STANDARD == "MA 4.1.1.e" & DOK %in% c(1, 1)	0	2	45	24	15
7	STANDARD == "MA 4.1.1.f" & DOK %in% c(1, 1)	0	3	54	29	18
8	STANDARD == "MA 4.1.1.g" & DOK %in% c(1, 1)	0	2	45	24	15
9	STANDARD == "MA 4.1.1.h" & DOK %in% c(1, 1)	2	3	54	29	18
10	STANDARD == "MA 4.1.1.k" & DOK %in% c(1, 2)	0	2	45	24	15
11	DOMAIN == "NR"	7	10	378	202	126
12	STANDARD == "MA 4.1.2.b" & DOK %in% c(1, 1)	0	2	54	29	18
13	STANDARD == "MA 4.1.2.c" & DOK %in% c(1, 1)	0	3	75	40	25
14	STANDARD == "MA 4.1.2.d" & DOK %in% c(1, 1)	0	2	54	29	18
15	STANDARD == "MA 4.1.2.f" & DOK %in% c(1, 1)	0	2	45	24	15
16	STANDARD == "MA 4.1.2.g" & DOK %in% c(1, 1)	0	2	45	24	15
17	DOMAIN == "NO"	8	10	273	146	91
18	UDOMAIN == "NUM"	17	17	651	348	217

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1560	834	520
19	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	2	17	131	72	45
20	STANDARD == "MA 4.2.1.a" & DOK %in% c(1, 2)	1	3	75	40	25
21	DOMAIN == "AR"	1	2	75	40	25
22	STANDARD == "MA 4.2.2.a" & DOK %in% c(1, 2)	3	5	90	48	30
23	DOMAIN == "AP"	3	4	90	48	30
24	STANDARD == "MA 4.2.3.a" & DOK %in% c(2, 2)	1	3	54	29	18
25	STANDARD == "MA 4.2.3.b" & DOK %in% c(2, 2)	1	4	90	48	30
26	DOMAIN == "AA"	4	5	144	77	48
27	UDOMAIN == "ALG"	10	10	309	165	103
28	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	2	10	62	34	21
29	STANDARD == "MA 4.3.1.b" & DOK %in% c(1, 2)	0	2	54	29	18
30	STANDARD == "MA 4.3.1.c" & DOK %in% c(1, 2)	0	2	54	29	18
31	STANDARD == "MA 4.3.1.d" & DOK %in% c(2, 3)	0	2	54	29	18
32	STANDARD == "MA 4.3.1.e" & DOK %in% c(1, 2)	0	2	45	24	15
33	STANDARD == "MA 4.3.1.f" & DOK %in% c(1, 2)	0	2	54	29	18
34	STANDARD == "MA 4.3.1.g" & DOK %in% c(1, 2)	0	2	45	24	15
35	STANDARD == "MA 4.3.1.h" & DOK %in% c(1, 2)	0	2	45	24	15
36	DOMAIN == "GC"	6	6	351	188	117
37	STANDARD == "MA 4.3.3.a" & DOK %in% c(1, 2)	0	2	54	29	18
38	STANDARD == "MA 4.3.3.c" & DOK %in% c(1, 1)	0	2	45	24	15
39	DOMAIN == "GM"	2	2	99	53	33
40	UDOMAIN == "GEO"	8	8	450	241	150
41	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	2	8	91	50	32
42	STANDARD == "MA 4.4.1.a" & DOK %in% c(2, 2)	0	4	75	40	25
43	DOMAIN == "DR"	2	4	75	40	25

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1560	834	520
44	STANDARD == "MA 4.4.2.a" & DOK %in% c(2, 2)	2	5	75	40	25
45	DOMAIN == "DA"	3	4	75	40	25
46	UDOMAIN == "DTA"	6	6	150	80	50
47	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	2	6	30	16	10
48	POINTS	49	53			

Table D.2
Summary of Grade 6 Item Bank Counts by Constraint (Parts 1 and 2 Combined)

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1539	825	513
1		41	41	1,539	825	513
2	ITEM_TYPE == "Polytomous"	8	12	312	164	101
3	STANDARD == "MA 6.1.1.a" & DOK %in% c(1, 2)	0	3	48	26	16
4	STANDARD == "MA 6.1.1.b" & DOK %in% c(1, 1)	0	2	45	24	15
5	STANDARD == "MA 6.1.1.c" & DOK %in% c(1, 2)	0	3	48	26	16
6	STANDARD == "MA 6.1.1.d" & DOK %in% c(1, 2)	0	2	45	24	15
7	STANDARD == "MA 6.1.1.g" & DOK %in% c(2, 2)	0	3	48	26	16
8	STANDARD == "MA 6.1.1.h" & DOK %in% c(1, 2)	0	2	45	24	15
9	STANDARD == "MA 6.1.1.i" & DOK %in% c(1, 1)	0	2	30	16	10
10	DOMAIN == "NR"	7	9	309	166	103
11	STANDARD == "MA 6.1.2.a" & DOK %in% c(1, 1)	0	2	45	24	15
12	STANDARD == "MA 6.1.2.c" & DOK %in% c(1, 1)	0	2	45	24	15
13	STANDARD == "MA 6.1.2.d" & DOK %in% c(1, 1)	0	2	45	24	15

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1539	825	513
14	STANDARD == "MA 6.1.2.e" & DOK %in% c(2, 2)	0	2	45	24	15
15	DOMAIN == "NO"	4	5	180	96	60
16	UDOMAIN == "NUM"	12	12	489	262	163
17	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	2	12	99	53	32
18	STANDARD == "MA 6.2.1.a" & DOK %in% c(1, 2)	2	2	45	24	15
19	DOMAIN == "AR"	2	2	45	24	15
20	STANDARD == "MA 6.2.2.a" & DOK %in% c(1, 1)	0	3	54	29	18
21	STANDARD == "MA 6.2.2.b" & DOK %in% c(1, 2)	0	2	45	24	15
22	STANDARD == "MA 6.2.2.c" & DOK %in% c(1, 1)	0	3	48	26	16
23	STANDARD == "MA 6.2.2.d" & DOK %in% c(1, 2)	0	3	48	26	16
24	STANDARD == "MA 6.2.2.e" & DOK %in% c(1, 1)	0	3	48	26	16
25	STANDARD == "MA 6.2.2.f" & DOK %in% c(2, 2)	0	2	45	24	15
26	STANDARD == "MA 6.2.2.g" & DOK %in% c(1, 2)	0	3	48	26	16
27	DOMAIN == "AP"	8	8	336	181	112
28	STANDARD == "MA 6.2.3.b" & DOK %in% c(2, 2)	0	3	48	26	16
29	STANDARD == "MA 6.2.3.c" & DOK %in% c(2, 2)	1	3	48	26	16
30	STANDARD == "MA 6.2.3.d" & DOK %in% c(2, 2)	0	3	60	32	20
31	DOMAIN == "AA"	4	4	156	84	52
32	UDOMAIN == "ALG"	14	14	537	289	179
33	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	2	14	110	57	35
34	STANDARD == "MA 6.3.1.a" & DOK %in% c(1, 2)	1	2	45	24	15
35	DOMAIN == "GC"	1	2	45	24	15
36	STANDARD == "MA 6.3.2.a" & DOK %in% c(1, 1)	0	2	45	24	15
37	STANDARD == "MA 6.3.2.c" & DOK %in% c(1, 2)	0	2	45	24	15
38	STANDARD == "MA 6.3.2.d" & DOK %in% c(2, 2)	0	2	45	24	15

Constraint ID	CONDITION	LB	UB	Frequency of Items by Bank		
				1539	825	513
39	DOMAIN == "GO"	3	5	135	72	45
40	STANDARD == "MA 6.3.3.a" & DOK %in% c(2, 2)	1	2	45	24	15
41	STANDARD == "MA 6.3.3.b" & DOK %in% c(2, 2)	0	2	30	16	10
42	STANDARD == "MA 6.3.3.c" & DOK %in% c(2, 2)	0	2	30	16	10
43	DOMAIN == "GM"	3	3	105	56	35
44	UDOMAIN == "GEO"	8	8	285	152	95
45	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	2	8	57	31	19
46	STANDARD == "MA 6.4.2.a" & DOK %in% c(2, 2)	1	4	60	32	20
47	STANDARD == "MA 6.4.2.b" & DOK %in% c(2, 3)	1	4	60	32	20
48	STANDARD == "MA 6.4.2.c" & DOK %in% c(1, 2)	1	4	60	32	20
49	STANDARD == "MA 6.4.2.d" & DOK %in% c(2, 3)	1	3	48	26	16
50	DOMAIN == "DA"	7	7	228	122	76
51	UDOMAIN == "DTA"	7	7	228	122	76
52	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	2	7	46	23	15
53	POINTS	49	53			