South Dakota
Science Assessment

2024–2025

Volume 7:
Item Bank Maintenance in Three-Dimensional Science Assessment Programs

---

south dakota
DEPARTMENT OF EDUCATION
Learning. Leadership. Service.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

This special study presents a comprehensive investigation into the item bank maintenance (IBM) strategies for three-dimensional science assessment programs, developed and implemented across a multi-state consortium. Drawing on a multi-state collaboration and a shared item bank, the study explores item maintenance test design, calibration methodology, item parameter drift, and predictive factors for item stability. The findings inform best practices for proactive IBM plan to ensure validity and reliability in adaptive testing environments.

1. INTRODUCTION

A calibrated item bank is an indispensable tool in a modern online testing program. In an online testing environment with machine or AI scored items, the use of banked item parameters allows for immediate score reporting. Immediate score reporting greatly enhances the usability of test scores. Remedial actions can be taken not only more quickly but scores also provide a more accurate reflection of what students know and can do at the time of reporting because the amount of learning (loss) between the time the students are tested and the time the test scores are reported is minimal. Second, pre-calibrated item pools are necessary for adaptive tests. The advantages of adaptive testing are well documented and include more precise estimates of student proficiency across the range of proficiency, a test experience that is better tailored to the student's performance level, the possibility of more fine-grained reporting for aggregate units compared to fixed forms because a larger sample of items is administered across students, and increasing test security because the set and sequence of items that constitute a test event is different across students. It is therefore not surprising that many state assessments have adopted adaptive testing programs relying on large pre-calibrated items banks.

While there are numerous advantages working with a calibrated item bank, a periodic evaluation of item content, formatting, statistics and IRT parameters is central to the validity of the inferences made from assessment results. The Standards for Educational and Psychological Testing recommend that "Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported" (AERA et al., 2014, p. 103, Standard 5.7). This study proposes a comprehensive item bank maintenance (IBM) plan. We propose that the operational items in a bank are periodically reviewed by content experts, and re-field-tested to obtain updated item statistics and item parameters. Administering operational items in field-test slots provides the opportunity to obtain updated IRT item parameters, using the same process as when the items were initially calibrated (i.e., a re-field-test: reFT).

Our bank maintenance approach can be contrasted with a reactive ('repair') model where items are re-field-tested only after being flagged for, for example, item drift. Just like how in motorcycle maintenance the oil change schedule is mileage-based, we propose a bank maintenance schedule where items are re-evaluated by content experts and parameters are updated before any issues with the item or its statistical performance become apparent.

Like any maintenance plan, there is an effort and cost associated with carrying out our proposed IBM plan, but these costs are small compared to the amount of effort to address problems with operational items in the middle of a testing window. In other words, repairs are more costly than maintenance. In addition, there are often no good psychometric solutions when an operational item

becomes problematic during the testing window; all our typical solutions suffer from major drawbacks. For example, dropping the item from scoring has implications for both measurement precision and test validity (i.e., the reduced test may not meet blueprint). Scoring a test with and without the item and assigning the maximal score is a business rule preferred by many stakeholders because it does not disadvantage the students who got the problematic item. However, in the context of measurement errors, this practice results in an overestimation of the actual proficiency and can be considered unfair for the student who did not get the item. In sum, an IBM plan will not only be cost-efficient but also safeguard the validity of the intended inferences made from test scores (ultimately, a lack of validity will result in costs incurred when test results are legally challenged).

We propose that parameters of re-field-test items are always updated so that they are based on the most recent data. Everything else being equal, the most recent data will always provide the best item parameter estimates. This principle underlies the philosophy of post-equating, which arguably may be the preferred equating method from a pure psychometric perspective. By updating item parameters in a pre-equated bank on a regular basis, we are true to the idea that item parameters may fluctuate over time while maintaining the advantages of a pre-calibrated item bank.

In this report, we describe how an IBM plan was implemented and piloted for a large item bank for three-dimensional science assessments that are shared by multiple states. In addition to just being a feasibility study, a research design was built into the pilot study so that we could investigate factors that potentially contributed to fluctuations of item parameters over time. For example, while we could have limited our selection of items selected for re-field-testing to older items only adhering to an item maintenance schedule primarily based on the age of the item, we purposefully also selected items that became part of the bank in the year after the Covid-19 pandemic closed schools and in more recent years.

In the following sections, we shortly describe the background of the three-dimensional science assessments, a spiraled test design, how item maintenance can become part of the item development and life cycle, and our approach to item calibration and detection of changes in estimated item parameters. Last, we explore factors that predict changes in item parameters and that could potentially be considered together with the mere age of items when establishing an item maintenance schedule (i.e., the schedule could be shorter for item types that tend to have item parameters that are more prone to change over time).

2. BACKGROUND

Cambium Assessment, Inc (CAI) collaborates with a group of states to build and maintain a large item bank for assessing three-dimensional science standards (3DSS). Each science standard is focused on the combination of a discipline core idea, a science and engineering practice, and a crosscutting concept. This distinct nature of 3DSS calls for a unique approach to the assessment development. CAI worked with a group of states to develop a common approach towards item development. All states share common item specifications and item development and review processes, resulting in a large, shared bank from which individual states build their own science assessments. All items are calibrated on a common scale.

Science assessments developed under the three-dimensional framework rely on multiple interaction items or 'clusters' centered around a common scientific phenomenon. In an item cluster, a student is presented with a real-word scenario related to a single performance expectation and is asked to interact with the presented stimulus material in various ways. For each item cluster, a series of explicit assertions can be made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses. Scoring assertions can be supported based on students' responses in one or more interactions within an item cluster. These assertions are binary (true/false) indicator variables and are the basic units of analysis in our three-dimensional science assessments. The IRT model that is used to calibrate the items on a common scale explicitly takes into account the local dependencies between assertions pertaining to the same cluster (see Section 7, *Calibration and Detecting Drift* of this report). In addition to item clusters, stand-alone items are developed to increase the number of standards that can be covered by a single test event. Stand-alone items are shorter and more traditional items, typically containing a single interaction. They are also scored with assertions, but the number of assertions in a stand-alone item typically ranges from one to three assertions, where an item cluster typically contains six or more assertions. A typical science assessment consists of six clusters and 12 standalone items for each student, though there are variations across states, subjects and grades.

3. SPIRALED TEST DESIGN

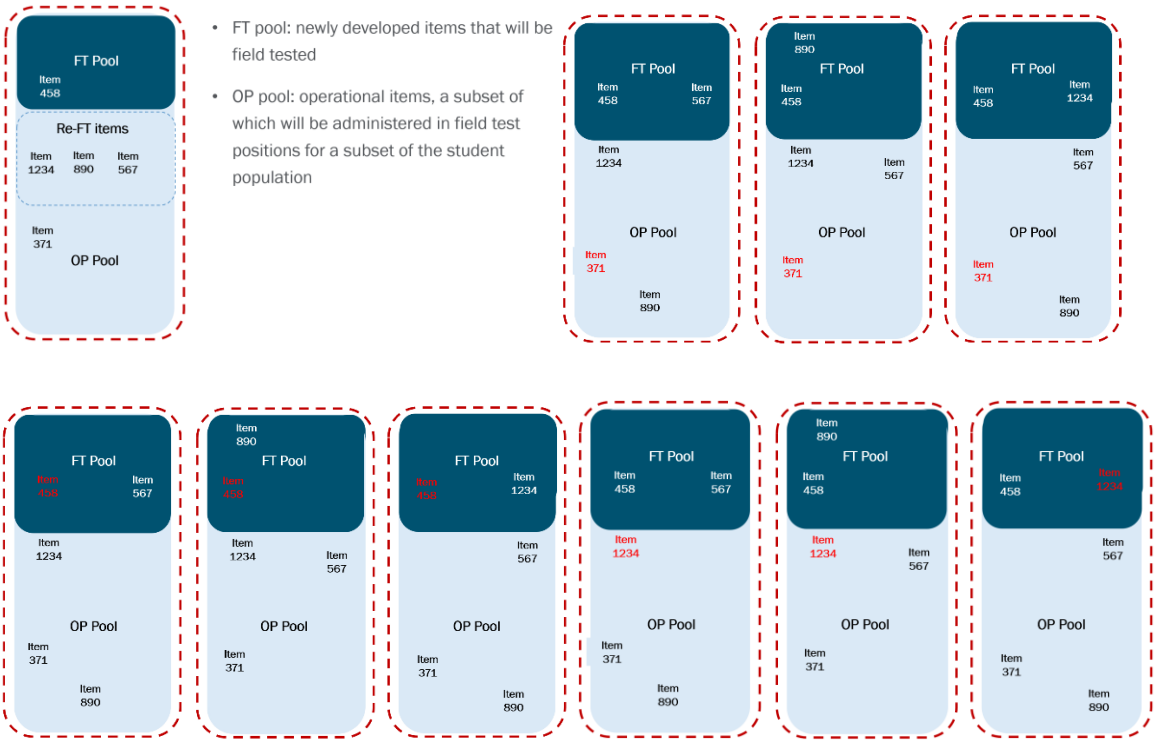
Most states implemented an adaptive test design. In most test designs, a student receives either one field-test cluster or four field-test stand-alone items in addition to scored operational items. Positions of field-test items are randomized across test events to average out item position effects.

In the simplest design, operational items due for a parameter refresh can be totally removed from the operational pool and added to the field test pool. This will ensure that item parameters are refreshed under the same conditions as new field-test items are calibrated. Note that in the first years of science assessments, a variety of test designs was used: (a) operational field-tests, (c) independent field-tests, (c) separate test segments with three-dimensional science items administered right before or after the legacy (pre-3DSS) science test, and (d) field-test items embedded within an operational test (the current test design). It is quite common for new testing programs to adhere to operational and independent field tests in the first year of an assessment to build out the item bank, and transition to an operational test with embedded field-test slots one the bank is sufficiently large. To the degree that the test design affects parameter estimates (e.g., because item positions tend not to be entirely random in an operational field-test due to the fact that they are selected to satisfy a test blueprint, or because student motivation is likely lower in an independent field-test and may affect different item types differently), a periodic update of item parameters is recommended especially for items that were calibrated in the early years of a testing program. Another reason for updating parameters for items calibrated in the beginning of a testing program is that instructional changes are more likely to take place following the adoption of new standards – presumably an intended consequence of adopting new standards – which may affect the psychometric properties of items.

While appealing in its conceptual simplicity, simply moving operational items to the field test pool comes with substantial limitations. First, the operational bank may be too shallow by having too many operational items re-field-tested for some elements of the blueprint. This will also result in high exposure rates for the remaining items in those set of items, which ultimately is a threat to test security. In more severe cases, removing too many items at once from the operational pool will result in test blueprint violations or reduced measurement precision due to less test adaptivity for certain regions of the proficiency scale. In general, it is desirable to keep the operational pool constant across years, or at the very least avoid shrinking the pool substantially. Related to the first point, it may be impossible to treat operational items as field-test items for accommodated forms because the item pool for those tests is typically smaller. This would result in a situation where the same item remains operational for accommodated test forms but becomes a 'field-test' item for regular forms. This is a situation we generally try to avoid.

As an alternative, we propose a test design that is based on the idea of spiraling fixed forms. Rather than using fixed forms, the idea of spiraling is applied to entire item pools. In the first step, the number of operational items that will be recalibrated is determined based on the IBM business rules (i.e., did an item reach its 'expiration date' so that it is due for maintenance) and the available field-test capacity. Second, based on an examination of the operational pool, we determined how many operational items at a time can be removed from the operational pool without jeopardizing the bank or the functionality of the adaptive item selection algorithm. Based on these two steps, the operational bank is divided into different versions or sets, where a subset of the items to be recalibrated is moved from the operational pool to the field test pool in each version. For example, if there are 30 items to be recalibrated out of 100 operational items, three versions of the test are created where each version of the operational bank consists of 90 items. Of these 90 items, 70 items are in common across all versions; these items are not recalibrated in the current year. The 20 remaining items are obtained by removing a different set of 10 operational items from each bank, and adding them to the corresponding field test pool (i.e., re-field-test items). During test administration, each student is randomly assigned to one of the three versions of the operational and corresponding field-test pool. This process ensures that all items in the field-test pool (newly developed field-test items or re-field-test items) are administered to a representative sample of the student population. Figure 1 illustrates the process of constructing the item pools under this spiraled test design. At the top left, the item bank is portioned into a field-test pool with newly developed items never field-tested before (e.g., Item 458) and an operational bank further partitioned into items that will be recalibrated (reFT items; Items 1234, 567, and 890) and operational items that will not be recalibrated (Item 371). At the top right, Item 371, which will not be recalibrated, is part of the operational pool for each version of the bank. Bottom left: Item 458, a newly developed item, is part of the field-test pool for each version of the bank. Bottom right: Item 1234, a re-field-test item, remains part of the operational pool in two versions, but is moved to the field-test pool in the third version of the bank. Similarly, Items 567 and 890 remain operational in two versions of the bank but become part of the field-test pool in the third version.

Figure 1. Illustration of a spiraled test design



As can be seen in Figure 1, as items are selected from the field-test pool with equal probabilities, the expected exposure of true field-test items is three times (in the case of three versions of the bank) as high as the expected exposure for re-field-test items. CAI modified its field-test item selection algorithm to incorporate field test weights that allows control the exposure of field-test items. In this case, this is accomplished by setting the relative weights for re-field-test items to three times the weight of true field-test items. In practice, the determination of field-test weights also takes into account the relative proportions of item clusters and stand-alone items.

4. Content Approach to Item Drift

In this section, an overview of the life cycle of an item is presented, and how the idea of item maintenance triggers an additional review step (i.e., items are not forever). It describes the content review component of that process, focusing on how science item content is checked for accuracy, relevance, and sensitivity over time as part of the larger item life cycle.

4.1 Background of the Three-Dimensional Item Bank

In 2017, a group of 12 states using three-dimensional science standards (3DSS) signed a Memorandum of Understanding (MOU) to form a collaborative group (hereafter referred to as "the MOU"). The MOU's ongoing goal is to develop a shared bank of items for summative science

assessments. The item bank includes content developed by Cambium Assessment, Inc. (CAI) and following states: Arkansas, Connecticut, Hawaii, Idaho, Indiana, Montana, New Hampshire, Oregon, Rhode Island, Utah, West Virginia, and Wyoming. These states collaborated with CAI to create item and test specifications aligned to 3DSS, reviewed by state experts and educators. Items are selected from the shared bank based on alignment with state blueprints and undergo further review before use.

4.2 ITEM LIFECYLE

The life cycle of a science item developed to align with the 3DSS at CAI begins with the conceptualization and design phase, during which content experts or participating state identify relevant scientific phenomena aligned to specific science standard performance expectations. These phenomena become the anchors for item clusters or stand-alone items that integrate three-dimensional science learning: Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs).

Next, CAI's item writers develop stimuli, interactions, and scoring assertions, which are reviewed internally for alignment, clarity, bias, fairness, and accessibility. States then engage in multiple review steps through Content Advisory Committees and Bias and Sensitivity Committees, ensuring each item meets both scientific and cultural standards. Some states host Item Writer Workshops to seed cluster development with real-world phenomena selected by local educators.

Following development, items undergo field testing, where they are administered to students—typically in two states, including the item's origin state. Typically, items appear as embedded field-test items in operational forms. Data and student responses from these administrations are reviewed during Rubric Validation (to confirm scoring rules) and Data Review (to examine psychometric properties). Only items that pass both steps are promoted to the shared operational item bank.

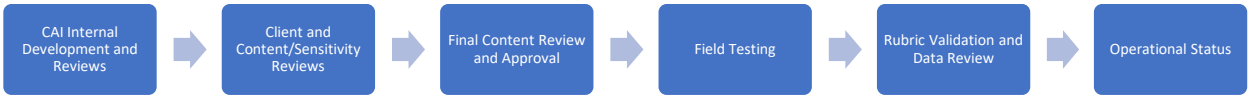
During operational use, items are selected by states according to their blueprint needs and used in adaptive test forms. Item parameters (e.g., difficulty, discrimination) are fixed using initial calibration and stored for scoring purposes.

Items undergo a content and sensitivity review to verify the scientific validity of the stimuli, ensure alignment with current events, and assess whether rendering or emotional content could affect student experience. Items passing this review are selected for re-field-testing (re-FT) using a spiraled test design, in which the same item is presented as operational for some students and field test for others, allowing for valid recalibration.

Recalibrated item parameters are compared to the original values to detect item parameter drift (IPD). Items that show notable drift are flagged for additional review to determine the source, such as exposure, shifts in curriculum, or changes in scientific familiarity. Finally, items are either retained, revised, or retired based on a combination of psychometric and content judgments.

This life cycle ensures items remain scientifically accurate, fair, and psychometrically sound for all participating states over time. Figure 2 shows a typical item lifecycle at CAI.

Figure 2. Item Lifecycle



4.3 Content Review and Maintenance Trigger

In CAI's approach to maintaining a valid and reliable standards-aligned item bank, the item bank maintenance (IBM) process serves as a proactive checkpoint in the item life cycle. IBM is triggered once an item reaches a designated "shelf life," typically seven or more years after its original operational use and calibration. This time-based threshold is deliberately chosen to preempt potential content or statistical degradation, mirroring preventive maintenance models in other domains.

The first step of IBM is a content-based re-evaluation, described in an earlier section, conducted by the content team. Items are reviewed for continued scientific accuracy, proper rendering, alignment to current instructional standards, and sensitivity to recent events or evolving cultural norms. No statistical analysis is involved—the goal is to ensure that the item is still appropriate to present to students in today's educational and social context. Items that fail this review are removed from the operational pool and either edited for future re-field-testing or permanently retired. Items that pass proceed to the next step: psychometric review and re-field-testing.

At this point, the psychometric team becomes involved, but not yet to evaluate item statistics—instead, they work with the content team to determine which items will be re-field-tested, considering several operational constraints. These include:

- The age of the item and time since its last calibration,
- The availability of field-testing capacity in participating states,
- Contractual obligations and developmental pipelines that may limit the number of items field-tested in a given year.

Once selected, the items are assigned to a spiraled test design, allowing them to be re-administered to a representative student sample. Only then does the psychometric team evaluate parameter stability, comparing the original IRT parameters to those estimated from the new administration. Items that show signs of item parameter drift (e.g., significant shifts in difficulty or discrimination) are flagged and referred back to the content team for root-cause analysis.

Thus, IBM fits between operational use and re-field-testing, acting as the formal trigger for deeper psychometric re-analysis. It ensures that items are not used indefinitely without scrutiny and enables proactive updates before any degradation in measurement validity becomes a problem in live scoring or reporting. This maintenance-first model reinforces test validity while minimizing disruptions to operational testing.

4.4 Content and Sensitivity Review Procedures

Eligible items undergo review for:

- Scientific accuracy and relevance of the stimulus materials
- Proper functionality of all item interactions
- Sensitivity issues in light of current events or social context

Special attention is paid to 3DSS items in Earth and Space Science, which often involve natural hazards or human impact. Even if scientifically accurate, items must not be emotionally distressing to students.

4.5 Examples of Content Drift and Sensitivity Concerns

Scientific developments and increased access to information may shift how familiar or novel a phenomenon feels to students. For example, topics like DNA electrophoresis—once advanced—are now common in high school labs, potentially reducing the difficulty of related items. Sensitivity reviews will verify that items are not biased, offensive, or emotionally upsetting to students. For example, the Earth and Space discipline of the 3DSS includes standards for natural hazards, severe storms, and human impacts on the environment at each grade level. While the content remains accurate, it is important for the student experience that items are free of sensitive or triggering topics. CAI has recent experience in these whole-bank sensitivity reviews. Such reviews were completed in response to the COVID-19 pandemic, the Maui wildfires, and Hurricanes Helene and Milton.

4.6 Conclusion

The content review component of IBM is a critical safeguard for ensuring that items remain scientifically accurate, contextually appropriate, and fair to all students. As part of a broader maintenance framework that includes psychometric recalibration, this process enables states to uphold high standards for validity and comparability across years. Ongoing improvements to this content review cycle—including criteria for sensitivity and real-world context—ensure that assessment content evolves alongside both the science and the students it aims to measure.

5. Item Selection and Design

The data collection and research design for bank maintenance must balance operational constraints for multiple statewide assessments while facilitating the study of item parameter drift. The states that participated in the item maintenance study chose to accept a subset of items from the Shared Bank to their item operational item pool based on state-specific blueprints. This leads to item pool differences across states, where an item may be accepted into one state's operational pool but not another's.

Selecting and assigning items for re-field-testing must follow a set of business rules: an item must be part of the state's operational pool, meanwhile the state must have sufficient field test capacity for both re-field-test items and true (newly developed) field-test items. In addition, to explore potential factors contributing to item parameter drift across years, we applied additional constraints, including the science discipline an item belongs to, the original year the item was calibrated, and item type (i.e., item cluster or standalone items). To be able to examine if item parameter changes

over time were substantially larger for a certain content area, we ensured a balanced distribution of re-FT items across science disciplines. The Shared Bank was initially calibrated in the 2018−2019 school year, with new field-test items added annually. To assess the impact of the calibration year, we also attempted to obtain an equal distribution of items based on their original calibration year. Likewise, to examine the effect of the state an item was calibrated in, we controlled whether items were re-field-tested in the same state.

−random selection process, while items were randomly selected from the Shared Bank and assigned to eligible states for re-FT. Then manual adjustments were made to meet operational requirements and balance the distribution of items across the disciplines, year of calibration, and item type. However, this hybrid approach was resource- and labor-intensive. In the second year (2024−2025), we implemented a mixed-integer programming (MIP) approach to optimize item selection. This paper detailed the methodology, constraints, and implementation of MIP in assembling re-FT forms across multiple states and results of item selection, highlighting the improvements made in the item bank maintenance (IBM) process.

5.1 MIXED-INTEGER PROGRAMMING FOR ITEM SELECTION

MIP is a mathematical optimization technique used in decision-making problems where some variables must be integers. MIP is widely applied for scheduling, logistics and resource allocation problems. In educational assessment, its use enables a more systematic and controlled approach to test form assembly (van der Linden & Diao, 2011). The MIP method consists of three key elements:

1. Objective Function: A function that needs to be maximized or minimized (e.g., maximum the number of re-FT items while not exceed the state's FT capacity).
2. Constraints: Rules that limit the solution space (e.g., state-specific capacity limitations).
3. Decision Variable: The set of variables that can be controlled.

Our MIP model was designed to assemble re-FT forms for operational assessments simultaneously across participated states. While selecting re-FT items for the 2024−2025 school year, there were a set of hard constraints that had to be met by any feasible solution, including:

1. State-Specific Pools: Items selected for re-FT in a state must be from that state's operational pool.
2. Cross-State Assignments: Each item must not be tested in no more than two states to maximize the number of re-FT items.
3. Capacity Limits: The number of items selected must not exceed state-specific FT capacity limits.

The objective function was defined by a set of soft constraints for research purposes, which were conditions that are preferred to be satisfied rather than strictly required. The violation on the soft constraints incurred a penalty, allowing the optimization process to find a balance between fulfilling competing soft constraints. For the Spring 2025 test administration, we included the following constraints:

1. Maximize the number of items selected and prefer assignment to two states. Due to the imbalanced field test capacity across participated states, it was not always possible to assign each item to exact two states. For make solutions feasible, we applied a two-step approach: items were firstly assigned to <= two states and then apply penalty to any items that had been assigned to only one state.
2. Maintain balance in item types (item clusters and standalones) across selected items.
3. Ensure each state receives a mix of cluster and standalone items.
4. Distribute items evenly across science disciplines to prevent concentration in a single content area.
5. Distribute items evenly across their original calibration year.
6. Aim for at least 30% of re-FT items to have been previously field-tested in the same state excluding the states that joined the MOU Shared Bank later than 2023, as they have field-tested so few items historically and it's very hard to meet the 30% requirement.

Since these constraints involved trade-offs, we assigned weights to indicate their relative importance. The optimization algorithm would prioritize constraints with higher weights while still considering those with lower weights. Finding the optimal set of weights for the soft constraints was a key step in our process. We experimented with multiple weight configurations, adjusting the relative importance of each constraint and observing its impact on the resulting item selection. Through iterative testing, we evaluated different weight combinations until we identified the final set the provided the most feasible solution, one that successfully balanced state capacities and item distribution. Table 1 below listed the final set of weights for each grade band.

Table 1. Soft Constraints and Relative Weights for the Objective Function

Soft Constraint	Weight
Maximize the number of items selected	8
Prefer to assign each item to two states	12
Ensure a balance of item types across selected items	8
Include both cluster and standalone items in each state	6
Maintain an approximately equal distribution of items across disciplines	2
Ensure a balance of distribution in calibration years (2019−2024)	4
When possible, aim to assign at least 30% of items that were previously field-tested in that state	2

5.2 RESULTS

The MIP model successfully optimized item selection for IBM while ensuring that all hard constraints were met, and soft constraints were balanced according to their assigned weights. Below, we presented the capacity utilization per state, along with the distribution of selected items by type, discipline, calibration year, and whether the 30% re-field-test rule was satisfied. Tables

2-4 below presented the total capacity available, the capacity used, and the distribution of items based on key characteristics. In our embedded field test design, students typically receive either one item cluster or four stand-alone items per state across three grade bands, including reFT items. When computing the total capacity, we use the pseudo-standalones while assuming one cluster is equivalent to four standalones. For example, a state has a capacity of 5 pseudo-standalones meaning we can field-test one cluster + one stand-alone item, or five stand-alone items in that state.

Across all states and grade bands, the MIP model respected state-specific item capacities. Most states had full capacity utilization, especially where the capacity was relatively smaller (e.g., SD elementary school: five available and 5 used). In cases with much higher capacity than others (e.g. IN), a small portion of the capacity remained unused because the algorithm failed to find a second state for selected items. The MIP method successfully selected a diverse mix of clusters and stand-alones per state, except UT tests that only used clusters. The optimization achieved near-equal representation across science disciplines for elementary and middle school. For high school, an equal distribution was not possible as some states only administered Life Science only tests. Across all grade bands, items were drawn from a spread of calibration years. The model achieved a reasonable balance, especially for high school. The 30% re-FT rule was not always met in several states, reflecting differences in historical item usage across states and the challenges in sourcing items while still satisfying other constraints with higher weights.

Table 2. Capacity Utilization and Item Selection Summary, Elementary School

State	Grade/Grade Band	Capacity Available	Capacity Used	Item Type (clusters/ standalones)	30% Re-FT Rule Met?	Calibration Year	Discipline
AR	3	23	23	5/3	-	2019: 27% 2021: 25% 2022: 13% 2023: 27% 2024: 4%	ESS*: 33% LS: 33% PS: 33%
	4	18	18	4/2	-		
	5	19	19	3/7	-		
CT	5	17	17	2/9	Met		
HI	5	8	8	2/0	Not Met		
ID	5	13	13	2/5	Met		
IN	4	37	37	4/18	-		
OR	5	28	28	6/4	Not Met		
RI	5	17	17	2/9	Met		
SD	5	5	5	1/1	Met		
UT*	4	16	16	4/-	-		
	5	12	12	3/-	-		

Note. UT science tests only included clusters. ESS: Earth and Space Science; LS: Life Science; PS: Physical Science.

Table 3. Capacity Utilization and Item Selection Summary, Middle School

State	Grade/Grade Band	Capacity Available	Capacity Used	Item Type (clusters/ standalones)	30% Re-FT Rule Met?	Calibration Year	Discipline
AR	6	25	25	6/1	-	2019: 25% 2021: 22% 2022: 24% 2023: 22% 2024: 6%	ESS: 32% LS: 33% PS: 35%
	7	30	30	5/10	-		
	8	17	17	4/1	-		
CT	8	27	27	5/7	Not Met		
ID	8	21	21	2/13	Met		
IN	6	64	59	10/19	-		
OR	8	17	17	3/5	Not Met		
RI	8	19	19	2/11	Met		
SD	8	5	5	1/1	Not Met		
UT	6	24	24	6/-	-		
	7	24	24	6/-	-		
	8	32	32	8/-	-		

Table 4.Capacity Utilization and Item Selection Summary, High School

State	Grade/Grade Band	Capacity Available	Capacity Used	Item Type (clusters/ standalones)	30% Re-FT Rule Met?	Calibration Year	Discipline
AR	10	22	22	5/2	-	2019: 25% 2021: 22% 2022: 24% 2023: 22% 2024: 6%	ESS: 17% LS: 58%* PS: 25%
CT	11	18	18	1/14	Not Met		
ID	11	10	10	1/6	Not Met		
IN	10	46	37	8/5	-		
OR	11	3	3	0/3	Not Met		
RI	11	6	6	½	Not Met		

Note. More LS items were selected for high school because some states had an LS-only test for high school.

5.3 DISCUSSION

The transition from a manual item selection in the 2023−2024 school year to a MIP-based approach in the second year yielded several notable improvements in efficiency, flexibility, and outcome quality. This shift not only streamlined the operational workflow but also enhanced the model's ability to meet complex constraints simultaneously.

One of the most immediate benefits of the MIP approach was the significant reduction in manual labor. In the first year, item selection required extensive human involvement, including iterative manual adjustments, repeated constraint checks, and time-consuming state-by-state coordination. In contrast, the MIP model automated the majority of these tasks, producing viable item sets in a fraction of the time. This automation greatly reduced the burden on psychometricians, allowing them to focus on reviewing and refining results rather than generating them from scratch.

The MIP framework was inherently more adaptable than a manual process. In the second year, the model was expanded to include both hard and soft constraints, each with assigned weights to reflect their relative importance. This weighting system enabled nuanced prioritization across multiple competing goals—such as discipline balance, item type ratios, re-FT rules, and calibration year distributions.

We were able to set on the constraint weights and fine-tune them until an acceptable solution was reached. In the second year, we were able to reach a more desirable item distribution for soft constraints. For example, the MIP-generated selections were more evenly distributed across years. In the first year, 44% of selected items were calibrated in 2019, 23% in 2021, and 28% in 2022 and only 5% were chosen from 2023. Those numbers range from 19% to 24% for the second year across 2019−2023.

In summary, the MIP-based item selection approach was faster, more consistent, and better able to meet a wide range of constraints—both operational and psychometric. It enhanced the overall quality of the item selection process while greatly reducing the human effort involved. Moreover, the MIP framework provides a flexible and scalable foundation for future work: as more states may participate, MOU owned items are involved, or other business rules may by applied, the model can be easily updated to reflect new hard or soft constraints. This adaptability ensures that future item selection cycles can respond quickly to changing business needs.

6. Calibration and Approach to Detecting Drift

Item parameter drift (IPD; Goldstein, 1993; Bock, et al., 1988) refers to a change in parameter values across testing administrations (i.e., over time). If an item bank is not monitored for drift over years, the percentage of drifting items and the magnitude of the drift may accumulate and negatively affect the measurement of the intended construct (Deng & Melican, 2010). One approach to drift studies involves recalibrating a subset of previously calibrated items to ensure the average drift across items is nondirectional and within bounds expected from sampling error (Wise & Kingsbury, 2000). This approach forms the basis for the item bank maintenance (IBM) pilot study in the 2023−2024 academic year.

CAI's science assessment calibration model supports the complexity of 3DSS-aligned items by modeling multiple scored assertions per item, accounting for local dependencies via nuisance dimensions, and calibrating across states using a multigroup Rasch Testlet Model. In our yearly calibration of the Shared Bank, the parameters of the operational items are fixed to their bank values, and the parameters of the field-test items and mean and variance of each group (state) are estimated using the marginal maximum likelihood (MML) method (Gibbons & Hedeker, 1992; Glas, et al., 2000). All students who attempted at least one field-test item are included in the calibration. For full model specifications and technical details, see Section 5, *Item Calibration*, in Volume 1 and its Appendix 1-C, *Calibration of the Shared Science Assessment Item Bank*.

Our research evaluated two calibration design approaches to shed light on the stability of the calibration methodology in the context of item bank maintenance: 1) concurrent calibration of true field-test (true FT) and re-field-test (reFT) items (i.e., operational IBM calibration approach) and

2) calibration of true field-test items only (i.e., yearly operational calibration approach prior to the IBM pilot study). After confirming the robustness of our operational IBM calibration method for true FT items, we then evaluated the magnitude and variability of parameter drift in the reFT items to assess both parameter and scale drift.

6.1 CALIBRATION DESIGNS

To ensure that the IBM design, described in Section 3 of this report, would not impact the calibration of true FT items, parameter estimates from two different calibration approaches were compared as part of the pilot study. In Approach 1, we calibrated true field-test and re-field-test items concurrently. Approach 2 was consistent with CAI's yearly calibration approach in which only true field-test items were calibrated; data from re-field-test items were excluded. In both approaches, field-test parameters were anchored on the bank parameters of operational items. Given the IBM spiraled test design, in which both the operational and re-field-test versions of the IBM items were administered during the school year, operational bank parameters of IBM items were included as anchors in both calibration approaches. The results presented below support the stability and robustness of true FT calibration results in the presence of re-field-test data.

Figure 3 presents elementary school (ES) and middle school (MS) assertion difficulty estimates of the true FT items for Approaches 1 and 2. The dashed diagonal line is the 45-degree line ($y = x + 0$), representing perfect agreement between the two calibration approaches. The largest absolute difference between the two approaches was 0.006 in ES and 0.002 in MS.

Figure 3. Assertion Difficulty, Cluster and Standalone Items

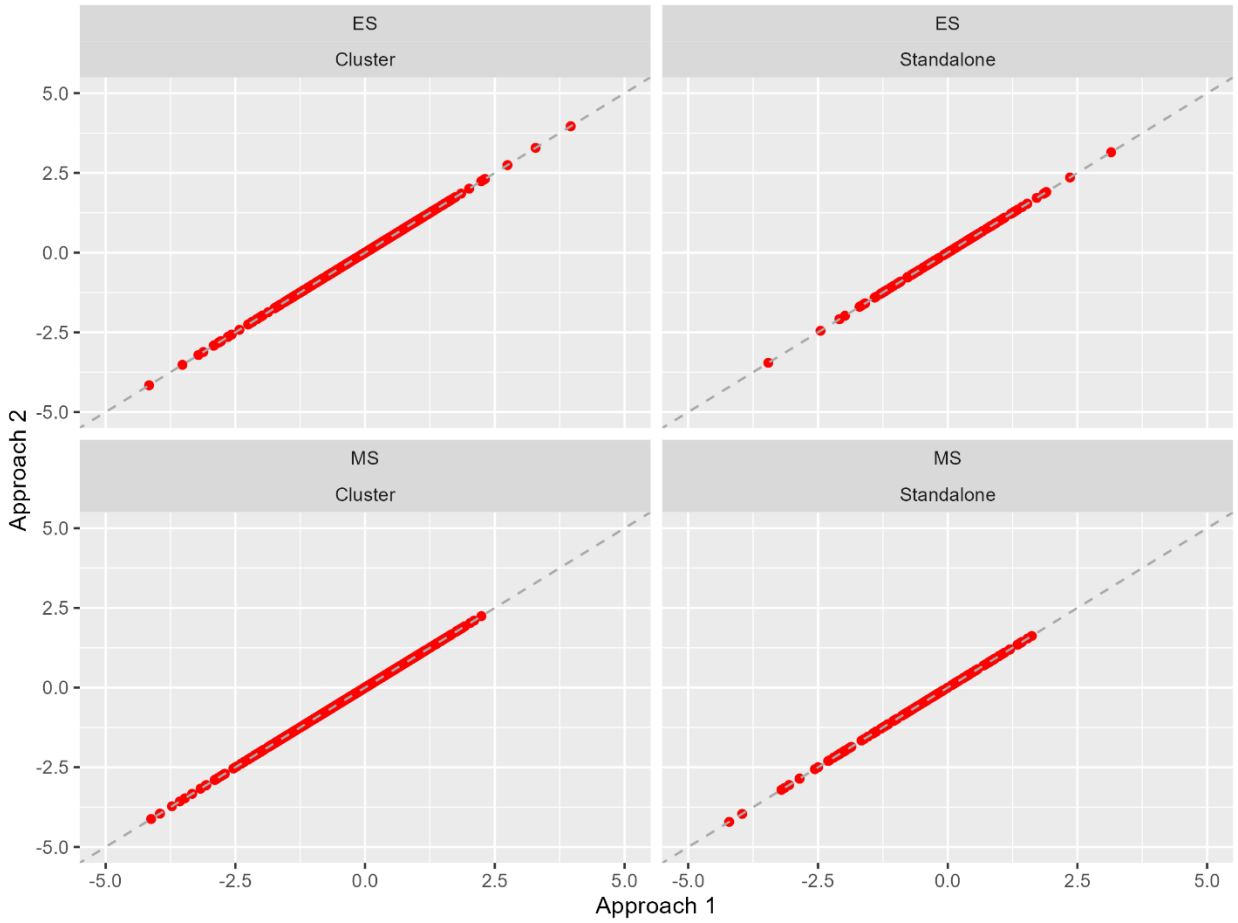


Figure 4 presents ES and MS estimates of the within-item local dependencies of the true FT items with four or more assertions. The largest absolute difference between Approach 1 and Approach 2 was 0.0001 in ES and 0.001 in MS.

Figure 4. Within-Item Local Dependencies

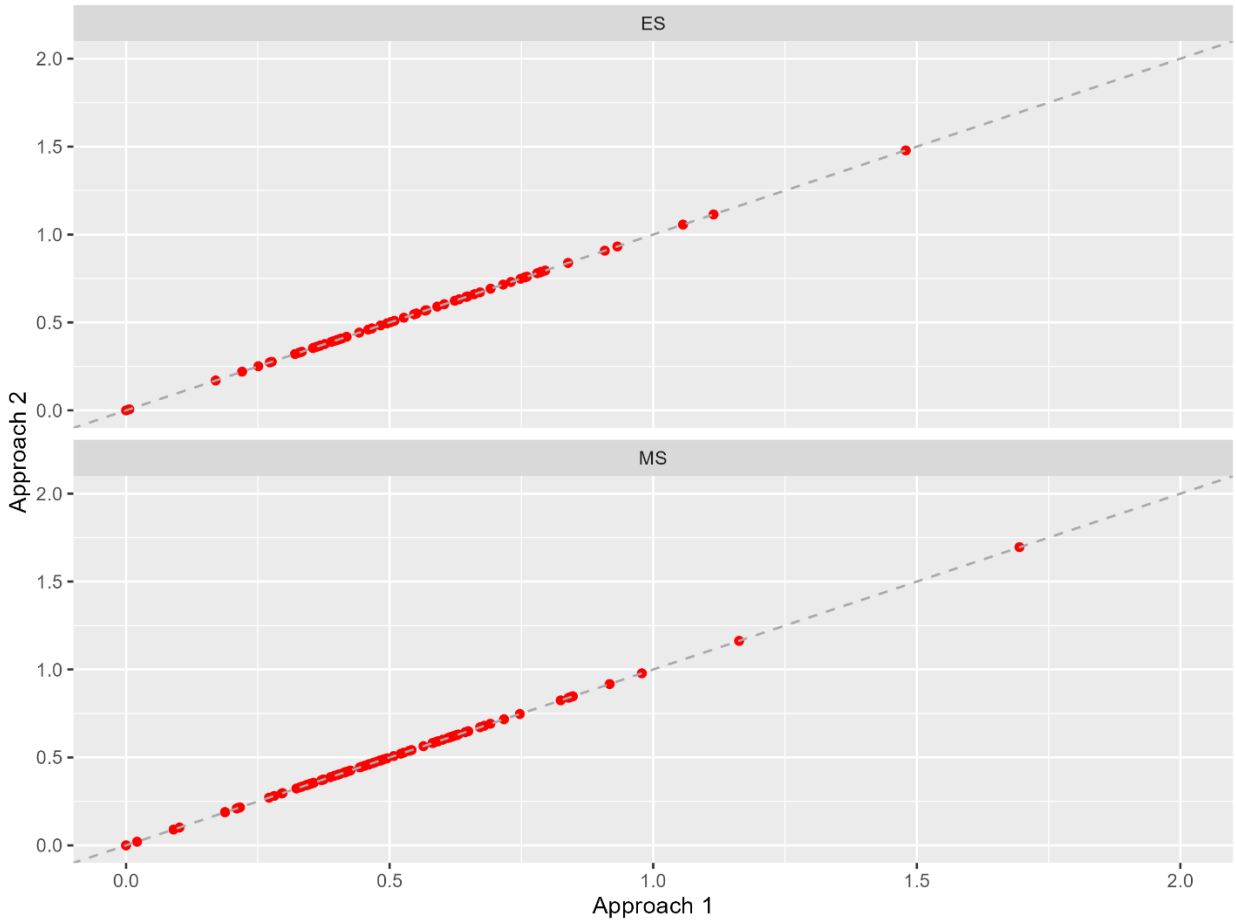
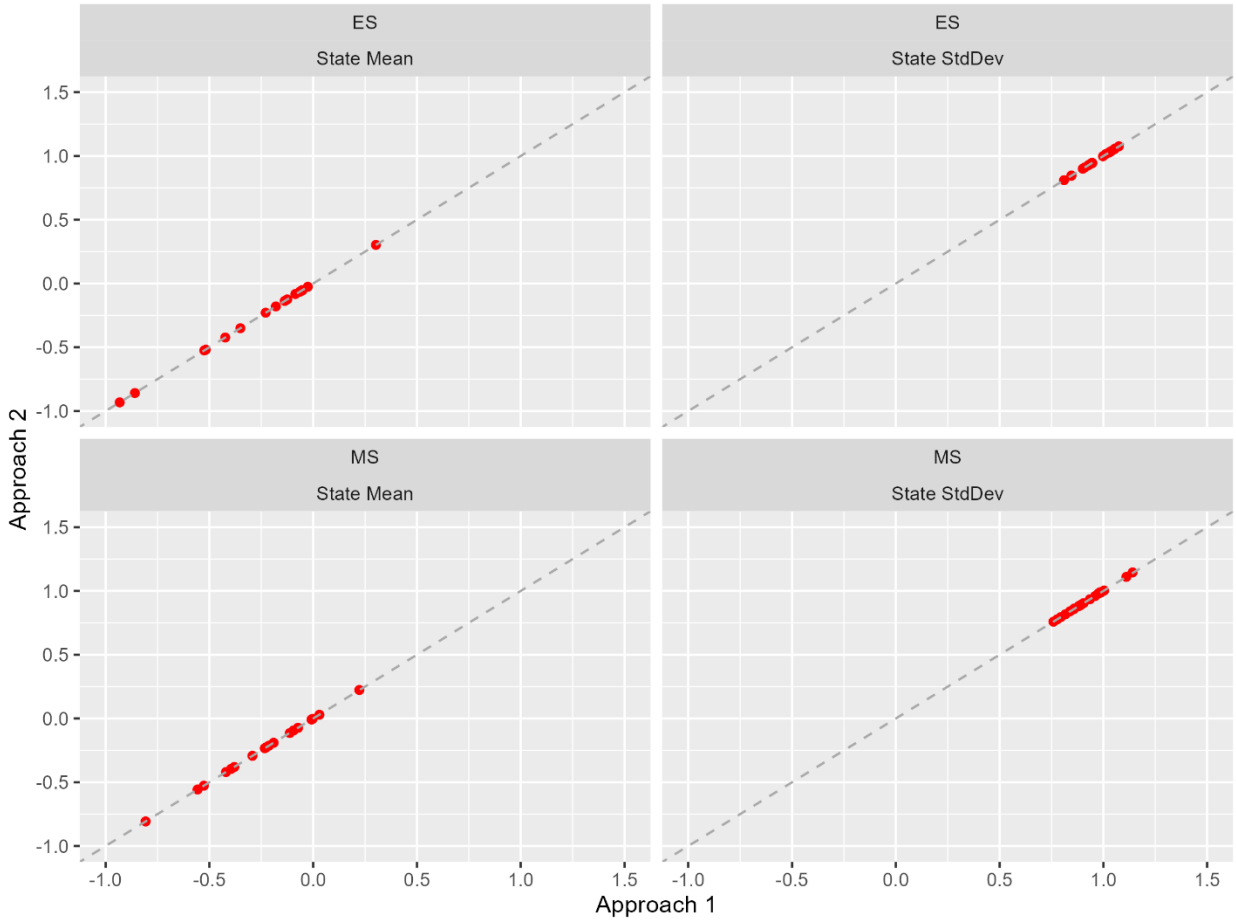


Figure 5 presents ES and MS group (state, US territory) means and standard deviations for Approaches 1 and 2. The largest absolute difference in group means between the two approaches was 0.003 in ES and 0.002 in MS. The largest absolute difference in standard deviations was 0.004 in ES and 0.006 in MS.

Figure 5. State and Territory Means and SDs



6.2 Recalibration Results from Spring 2024 Pilot

We recalibrated the re-FT items (Approach 1) and compared the bank values from the original calibration to the recalibration results to assess item and scale drift. We applied a simple measure of the difference in assertion difficulties obtained from the two separate calibrations. Scale drift can be inferred from the average difference between the old and latest location parameters. A consistent positive or negative difference indicates that the probability of a correct response, conditional on student ability, has changed systematically over time.

6.2.1 Bank and Recalibration Comparison

In elementary school (ES), the correlations between bank difficulties and recalibrated difficulties were r = 0.980 for clusters and r = 0.992 for standalone items. In middle school (MS), the correlations were r = 0.987 for clusters and r = 0.982 for standalones. These correlations show high stability in parameters between their banked values and re-field-testing. The bank and re-field-test difficulties are shown in Figure 6 for ES and Figure 7 for MS. Assertions with absolute differences larger than 0.3 are indicated in red.

Table 5 summarizes the raw differences between the bank and recalibrated difficulties. Raw differences take into account the direction of the change, with positive differences when the re-field-test difficulty is greater than the bank difficulty. We report both the mean and median differences by grade band and by item type; however, the median is more informative than the mean due to outliers in the data.

Within grade band, ES standalones show a slightly higher median shift than ES clusters (0.034 for standalones, 0.007 for clusters) while the median shift was slightly higher for MS clusters than MS standalones (-0.010 for clusters, -0.001 for standalones). Across grade bands, the overall raw median differences were negligible (0.010 in ES, -0.007 in MS). Because there was no systematic shift in difficulties in one direction or the other, we conclude that there is no evidence of scale drift.

Table 5. Bank and Recalibration Comparison, Raw Difference

Grade Band	Mean	Std. Dev.	SE (Mean)	Median	SE (Median)
ES – Cluster	0.047	0.246	0.018	0.007	0.023
ES – Standalone	0.005	0.141	0.023	0.034	0.029
ES – All assertions	**0.040**	**0.232**	**0.016**	**0.010**	**0.019**
MS – Cluster	-0.002	0.159	0.011	-0.010	0.014
MS – Standalone	-0.033	0.204	0.032	-0.001	0.040
MS – All assertions	**-0.007**	**0.168**	**0.011**	**-0.007**	**0.014**

Table 6 summarizes the absolute differences between the bank and recalibrated difficulties by grade band and by item type. The absolute differences reflect the amount of shift, regardless of direction. Within grade band, ES clusters show a slightly higher median shift than ES standalones (0.134 for clusters, 0.088 for standalones) while the median shift was higher for MS standalones than MS clusters (0.130 for standalones, 0.088 for clusters). Across grade bands, the median absolute difference was higher for ES items than MS items (0.121 for ES, 0.095 for MS). As shown in Figures 6 and 7, ES items had proportionally more assertions with shifts greater than 0.30 from their original banked values, indicating that they were relatively less stable than MS items. Similarly, the standard deviation of the mean is higher for ES than MS (0.165 for ES, 0.117 for MS), reinforcing that ES items exhibited more variability in parameter changes. These findings emphasize the need for continued monitoring and recalibration to ensure consistency in item performance across grade levels.

Table 6. Bank and Recalibration Comparison, Absolute Difference

Grade Band	Mean	Std. Dev.	SE (Mean)	Median	SE (Median)
ES – Cluster	0.179	0.174	0.013	0.134	0.016
ES – Standalone	0.111	0.085	0.014	0.088	0.017
ES – All assertions	**0.168**	**0.165**	**0.011**	**0.121**	**0.014**
MS – Cluster	0.113	0.112	0.008	0.088	0.010
MS – Standalone	0.152	0.137	0.022	0.130	0.027
MS – All assertions	**0.120**	**0.117**	**0.008**	**0.093**	**0.010**

Figure 6. Spring 2024: Pilot – Bank and Recalibration Comparison, Elementary School

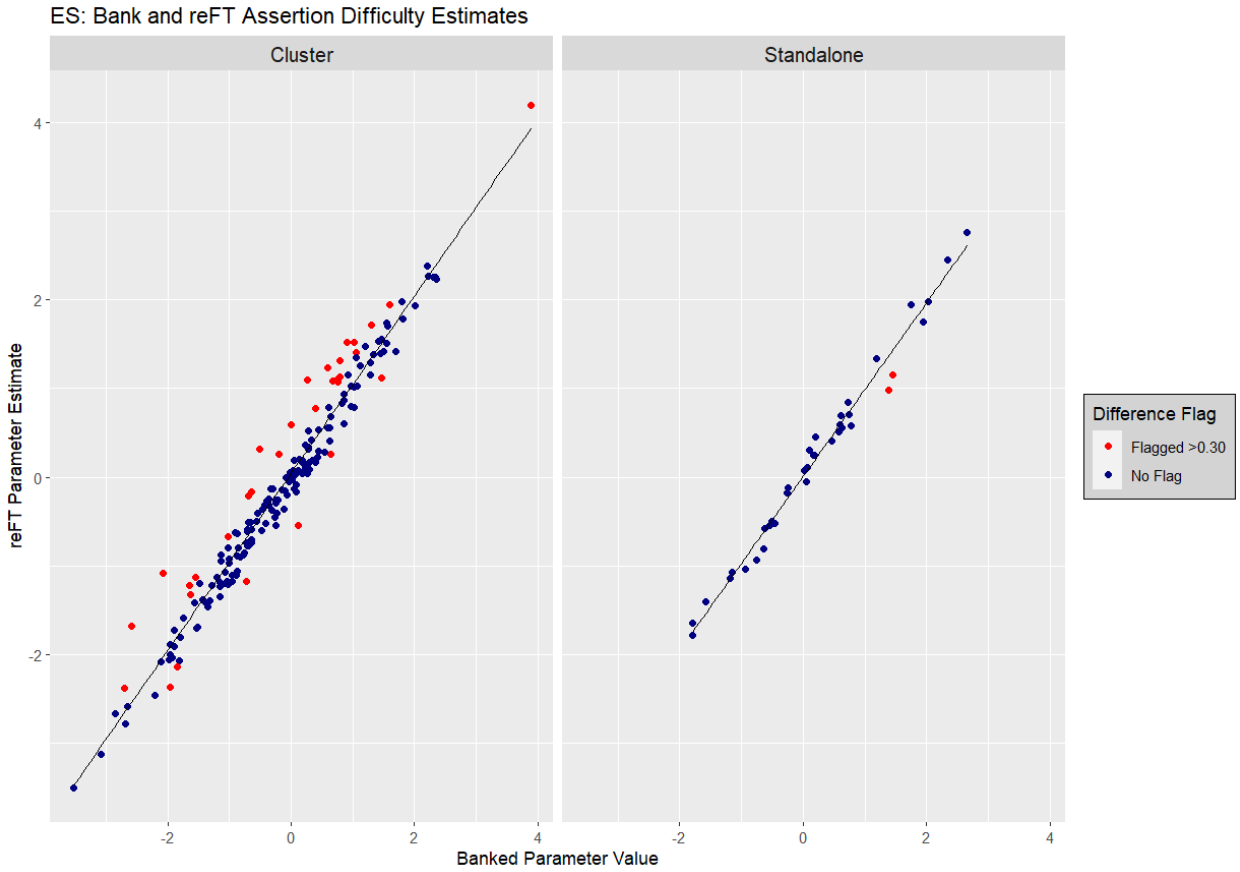
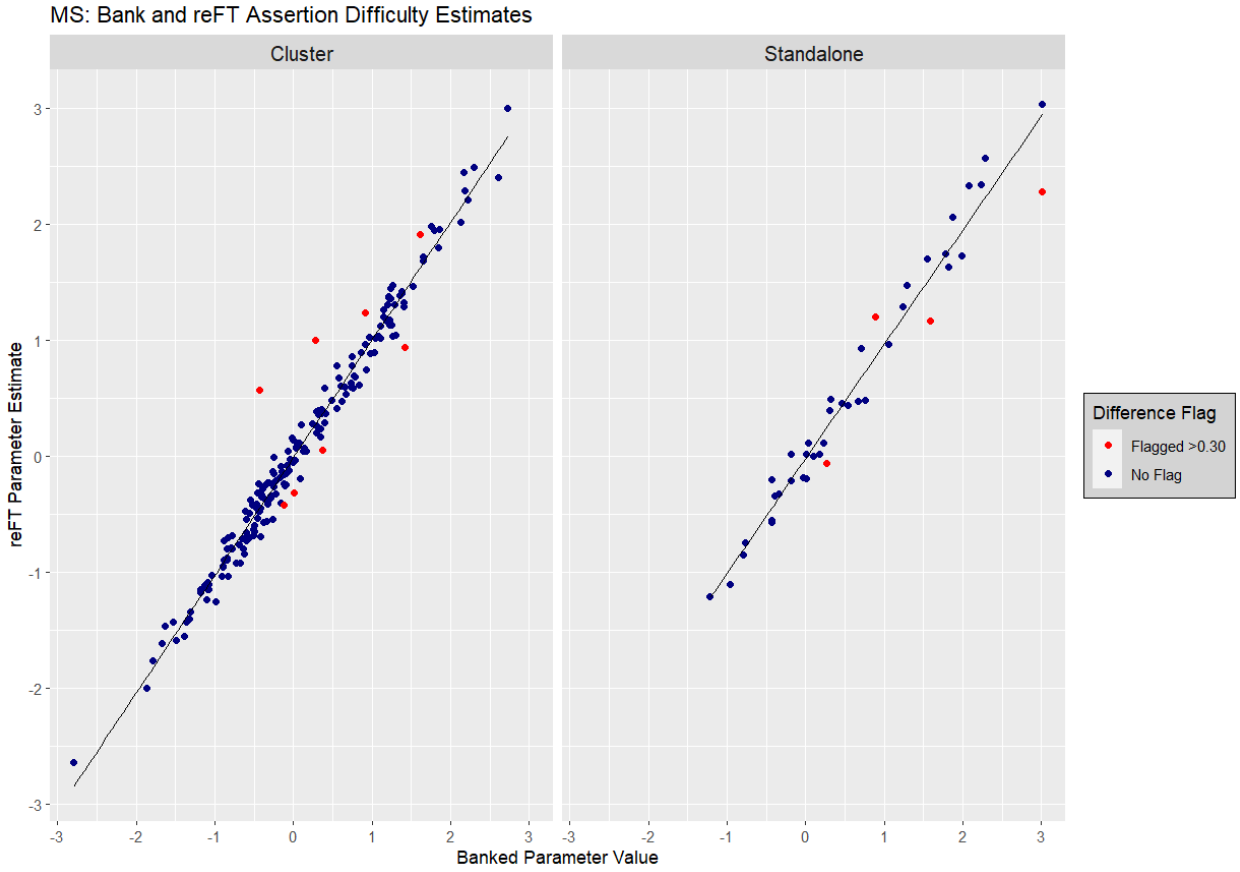


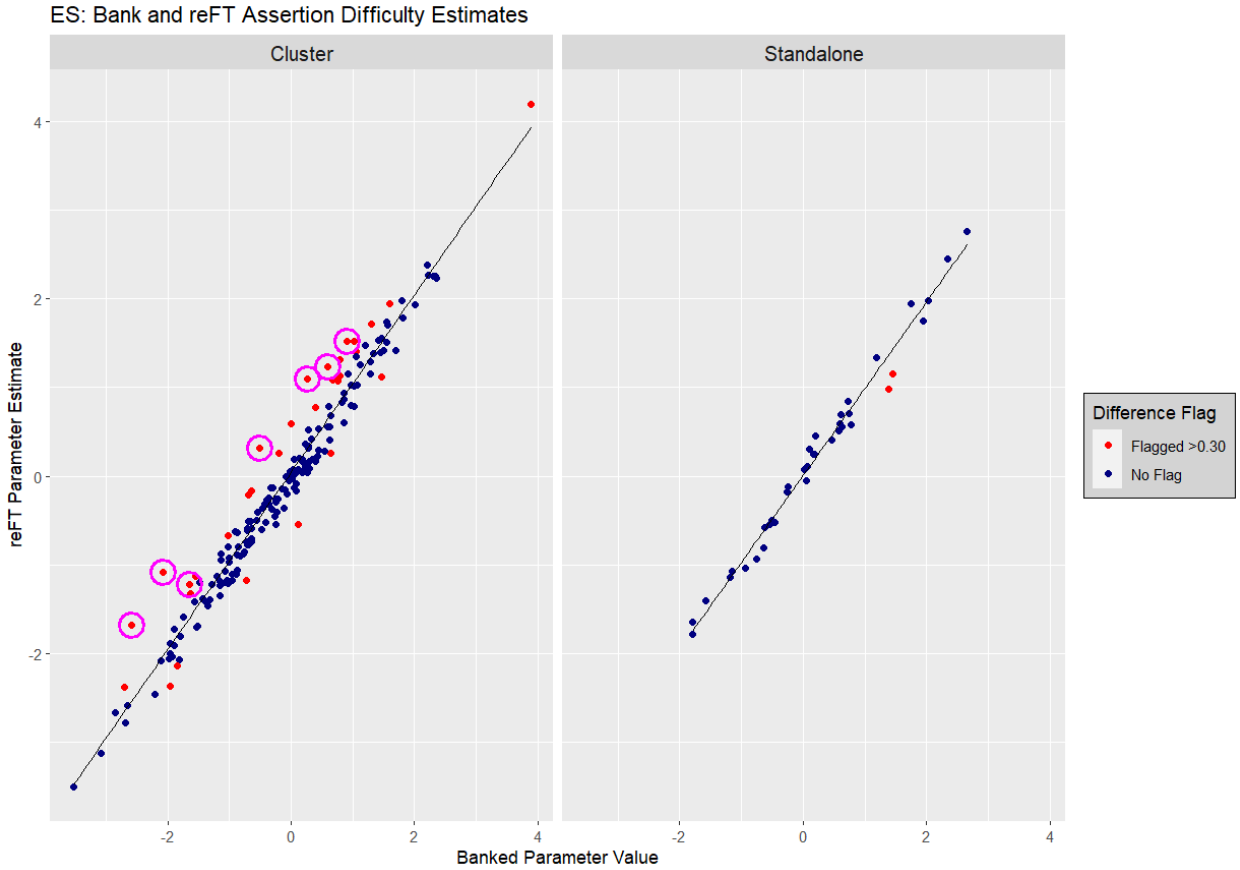
Figure 7. Spring 2024: Pilot – Bank and Recalibration Comparison, Middle School



6.2.2 Outlier Analysis

As noted above, parameters from the original calibrations were, in general, very stable, although there were assertions with drift greater than a magnitude of 0.3, especially in ES. Of the ten assertions with the largest absolute changes in difficulty, ranging from 0.622 to 1.000, over half belong to a single ES cluster item (shown in Figure 8 as pink circles). The item was flagged in one state for a rendering issue on Chromebook. While the rendering issue may not explain drift, this item has been rejected in all banks, beginning in 2024−2025.

Figure 8. Example of an Item Cluster with Drifted Assertions



6.3 Discussion

In Spring 2024, CAI successfully re-field-tested and recalibrated 98 items without impacting the true field-test item calibration. Overall, we found that the parameters and scale are both highly stable. However, elementary school items exhibited more fluctuation and variability in difficulty shifts than middle school items.

These findings underscore the need for ongoing recalibration and monitoring of item parameters to ensure the accuracy and fairness of the assessments. As new field-test items are introduced into the item bank, it will be important to continue recalibrating older items to mitigate the risk of drift accumulating over time.

7. Exploring Factors Related to Item Parameter Drift

While re-field-testing of the selected three-dimensional science standards (3DSS) items confirmed that assertion parameters remained largely stable, some degree of variation in parameters between different test administrations over time is inevitable. Differences in assertion or item parameter estimates between the original field-testing and re-field-testing highlight the need to understand the underlying factors contributing to these changes, as the purpose of item bank maintenance (IBM) is to identify and address such items before a significant drift happens. This paper

investigates predictors of item parameter changes using 3DSS test data with two primary objectives: (i) to examine the relationships between item parameter shifts and potential predictors, even when changes are small, to explain why such changes occur, and (ii) to identify key variables that should inform future item selection for re-field-testing, beyond simply considering item age when establishing an item maintenance schedule. Individual item's maintenance schedule can be expedited based on item characteristics that are more susceptible to parameter drift over time or item statistics that indicate parameter drift.

7.1 LITERATURE REVIEW

One of the most significant drivers of IPD is changes in educational curricula and societal knowledge. Goldstein (1983) and Bock, Muraki, and Pfeiffenberger (1988) discussed how shifts in instructional practices, technological advancements, and cultural developments can render test content obsolete.

Item exposure over time is also recognized as another factor influencing IPD (Bock, Muraki, & Pfeiffenberger, 1988; DeMars, 2004). Using simulation, Veerkamp and Glas (2000) showed that item difficulty decreases with increased exposure, with increased chance of correct responses. Conversely, item drift analysis has been used to assess item exposure (e.g., Giordano, Subhiyah, & Hess, 2005). Guo, Robin, and Dorans (2017) emphasized that IPD does not always follow a consistent pattern—some items drift gradually, while others can change rapidly. For example, items that become overexposed over time or are affected by gradual curricular change, may exhibit slow, progressive drift. In contrast, compromised items show a rapid change in parameters.

Sampling variability is another contributor to IPD, particularly when sample sizes are small. Swaminathan and Gifford (1983) noted that limited sample sizes can lead to unstable item parameter estimates, introducing artificial fluctuations. Wyse and Babcock (2016) explored how calibration timing and seasonal variations impact item parameter stability.

If there are changes between test administrations in things that can affect item parameters, such as item order or context (Yen, 1980), item position (Schnipke & Scrams, 1997; Wise & DeMars, 2006), multidimensionality of tests (Bejar, 1980), or estimation algorithm (Ban et al., 2000), they may contribute to item parameter change. Using the 3DSS assessment data, Cui (2023) found that items flagged "got harder" were positively associated with the difference between their average item positions in operational data and average item positions in original calibration data; conversely, item flagged as "got easier" were associated with earlier average positions.

Previous studies on item position and item exposure have investigated response time as data in addition to response accuracy. Research on item position showed that students showed higher rates of very short response time or item skipping toward the end of test (e.g., Akyol et al. 2021; Kroehne, Deribo, & Goldhammer, 2020). Studies on item overexposure have utilized response time as data to identify exposed items (e.g., Choe, Zhang, & Chang, 2018; van der Linden & Guo, 2008; Qian, Reckase, & Woo, 2016).

To detect IPD or to monitor operational item pools without re-field-testing, item residual analyses have been used as one method (Cui, 2023; Hambleton et al., 1991, pp. 59–61; Robin, Steffen & Liang, 2014). For the 3DSS assessments, item residuals (between the observed and the predicted scores based on IRT models) are routinely monitored to detect potential item parameter drift while items are used as operational scored items.

7.2 METHOD

7.2.1 Data

The re-field-testing of 98 selected items was conducted in the year 2023−2024.

7.2.2 Measure of Item Parameter Change (Outcome Variable)

Stand-alone items and item clusters in 3DSS assessments include one or more assertions, which have their own IRT difficulty parameters. We defined item-level parameter change as the average of absolute values of changes in assertion difficulties within an item. Figure 9 shows how the measure of item-level parameter change would be calculated for an imaginary item with six assertions. By taking absolute values, it captures both directions of parameter change without opposite directions of drift cancelling each other out. As this item parameter change measure can be only positive value, it only reflects the magnitude of change, ignoring directions.

Figure 9. Measure of Item-Level Parameter Change

reFT param	FT param	Difference (reFT - FT)	Absolute Difference	Avg Abs Diff
0.3	0.3	0	0	0.02
0.25	0.3	-0.05	0.05	
0.1	0.1	0	0	
0.45	0.4	0.05	0.05	
0.6	0.6	0	0	

7.2.3 Predictor Variables

As predictors for item parameter change, we adopted factors from the research design of the bank maintenance pilot study: calibration year (converted to item age), science discipline, and FT design. We also included predictors that were investigated in previous IPD research and related variables such as item position, item residual, and response time. Additionally, since the stability of item difficulty parameter estimates is associated with the difficulties themselves, (i.e., low/high difficulty items tend to have higher SEs), the original bank parameters were included. Although frequently discussed in the literature, item exposure was not included as a predictor, as exposure rates have been consistently decreasing over time due to the continuous expansion of item banks in the states making use of the shared item bank.

Item age. Item's age in years, e.g., an item that was originally field-tested and calibrated in Spring 2023 would have an age of 1 year in Spring 2024.

Science discipline. Three disciplinary areas of physical science, life science, and earth and space science. In regression analysis, earth and space science was used as a reference category for dummy coding.

Original FT design. For each item, the percentage of students across the four field-testing designs (shown in Table 8). For example, one item that was field-tested and calibrated in Spring 2019 had 24% LT, 0% IFT, 76% OFT, and 0% OP. The distribution of these designs varied by

year (See table 1) and therefore FT designs and item age were associated. FT design was also related to item positioning.

Table 7. Field-Testing Designs

Design	Description	Years (Spring)
LT	Legacy Test– 3DSS field-test items were administered in a separate segment before or after legacy (pre-3DSS) science items.	2019, 2021
IFT	Independent Field Test - Entire tests comprised of 3DSS field-test items presented in random positions.	2021
OFT	Operational Field Test - 3DSS field-test items were administered at random positions. Operational items were administered linearly on the fly and recalibrated after test administration; the recalibrated parameters were stored as bank parameters.	2019
OP	Operational Test with Embedded Field Test – 3DSS field-test items were administered at random positions. OP items were administered adaptively.	2021, 2022, 2023

Bank item difficulty. We used the absolute value of difficulty parameters from the calibration as both lower and higher bank item difficulties are expected to exhibit greater drift. Since the outcome measure in the study only reflects the magnitude of drift, absolute values were appropriate.

Relative item position in original calibration year. Item positions in pre-re-FT year (2023) or re-FT year (2024) were randomized and therefore only item positions in original calibration data were examined. Item's relative item position was calculated by dividing item position with test length. After that, it was centered around 0.5, which made the relative item position = 0 if the item was positioned in the middle of a test. Relative positions are averaged across all students and converted into absolute values.

Item residuals. First, assertion residuals for each student were calculated as the difference between the observed response and the expected probability of correct responses under the Rasch Testlet Model (for details, see Cui, 2023). Assertion residuals were then calculated by averaging assertion residuals across students. Next, to be consistent with the calculation of item parameter change, item residuals were obtained by averaging absolute assertion residuals within each item. Item residuals were calculated using pre-re-FT year (2023) data, where 93 out of the 98 items were administered as operational items.

Changes in response time. Average item's 80th percentile of response time (RT) in pre-re-FT (2023) minus average item's 80th percentile of RT in original calibration years, in minutes.

Item residuals and changes in RT were calculated using pre-re-FT (2023) data, the most recent year before the re-field-testing in 2024, as in a practical setting the most up-to-date item statistics would be used. In the pre-re-FT year, 93 of the 98 re-field-tested items were administered as operational items in adaptive test settings, with the other four items undergoing their initial field-testing.

7.2.4 Analysis

The bivariate relationships between item parameter change and predictors were examined using scatterplots, correlations, and regression (for science discipline and original FT design). Based on

the results of bivariate relationships, predictors that were related to item parameter change were included together in a multiple regression model.

7.3 RESULTS

Table 8 includes correlations between item parameter change and predictors (where correlations were applicable), and Figure 10 provides the visual illustration of the relationships.

Table 8. Correlations among variables

	V1	V2	V3	V4	V5	V6
(V1) Item Parameter Change	1	.205*	.299*	.354*	.413*	.130
(V2) Item Age		1	-.087	.366*	.226*	.150
(V3) Bank parameter			1	.029	-.005	-.240*
(V4) Item Position				1	.289*	.320*
(V5) Item Residual, pre-re-FT					1	.200
(V6) RT change, pre-re-FT						1

*p<.05

7.3.1 Bivariate Relationship between Item Parameter Change and Predictors

Item age. The older items were, the larger item parameter changes showed, with r=.205.

Science discipline. Even though some physical science items showed large parameter changes, a regression of item parameter change on science disciplines did not show significant regression coefficients or R^2 (=.041).

Bank item difficulty. Absolute bank parameters had a significant and positive correlation with item parameter change, with r=.299.

Relative item position in original calibration year. Absolute relative item positions in original field-testing had a significant and positive correlation with item parameter change, with r=.354, suggesting deviating from the middle of a test was associated with item parameter changes. A supplemental correlation analysis was conducted using item parameter drift defined as the average (non-absolute) changes in assertion difficulties, which reflects the directions of drift. The directional item parameter drift showed a correlation of r=-.423 with raw relative item position (non-absolute and non-centered), suggesting that items seen earlier in their calibration years were associated with positive drifts, i.e., item becoming more difficult, and vice versa.

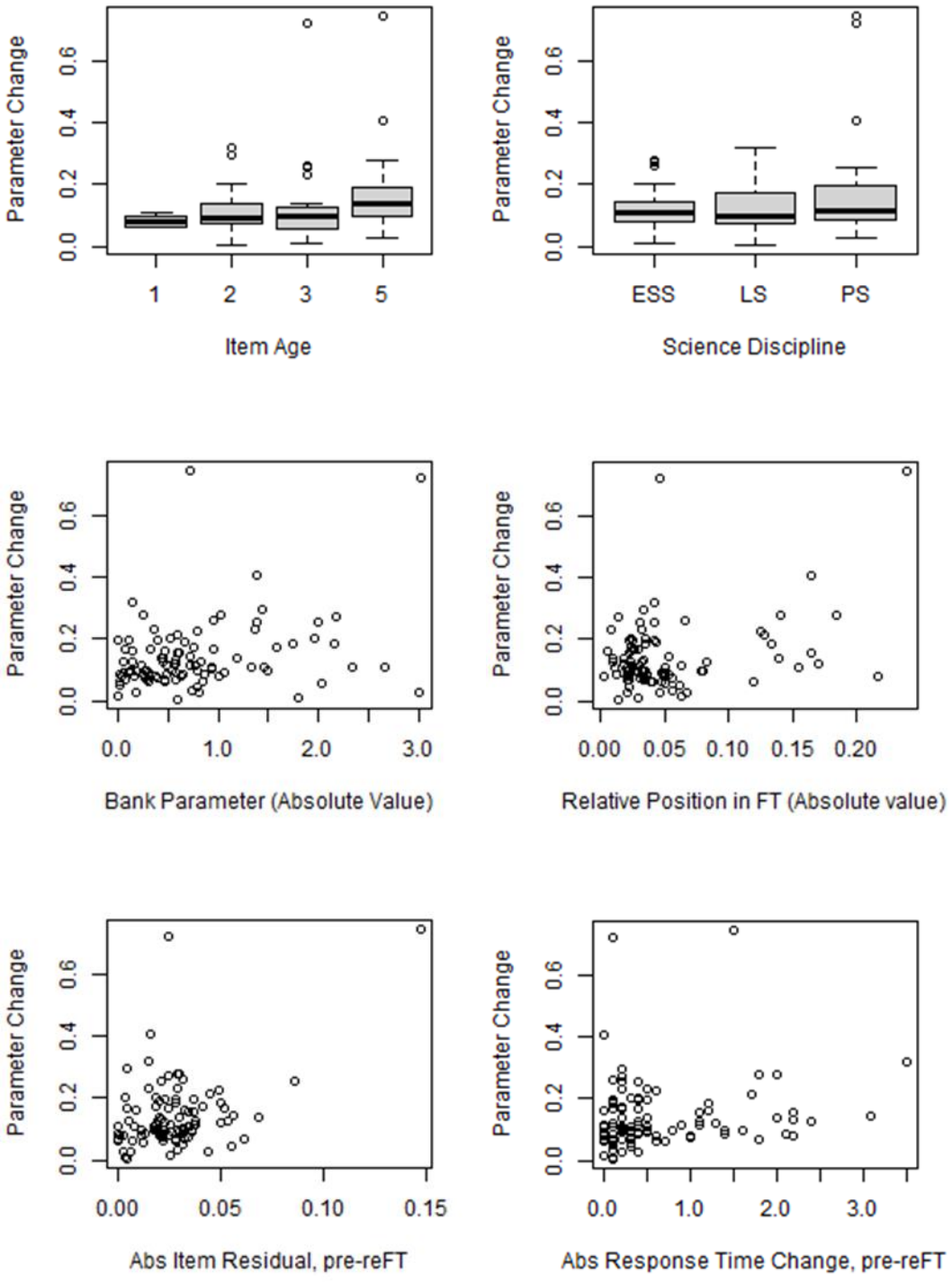
Original FT design. A regression model on FT designs showed an R^2 of .080 (p=.051). When item age and item position, which were confounded with FT designs, were controlled for, FT designs did not explain item parameter change (results are not presented in paper). In the regression, percentages of LT, IFT, and OFT were included as predictors, as OP percentage is linearly dependent on the other three (the sum of all percentages equals 100%).

Item residuals. The correlation between item residuals from pre-re-FT data and item parameter change was moderate (r=.413). While larger item residuals are associated with greater drift, the relationship was far from definitive, which highlights the importance of item bank

maintenance using re-field-testing. A supplemental correlation analysis between directional item parameter drift and item residuals (non-absolute) showed a correlation of r=-.543, suggesting positive residuals associated with negative drift, and vice versa.

Changes in response time. Response time (RT) changes between pre-re-FT and calibration data showed a correlation of .130 with item parameter change, and it was not statistically significant. A supplemental correlation analysis between directional item parameter drift and RT change (non-absolute) showed a correlation of r=-.210, implying that shorter response times, compared to calibration data, tended to be associated with item becoming more difficult, with a weak degree of association.

Figure 10. Visual Investigation of Predictors with Item Parameter Change



7.3.2 Multiple Regression with Selected Predictors

Among all predictors, item age, bank difficulty parameters, relative item position in calibration year, item residual from pre-re-FT data were selected for the following multiple regression analysis, to examine their effects on item parameter change altogether.

Table 9 shows the results of multiple regression of item parameter changes on the selected predictors. The regression coefficient of item residual was 1.890, which suggests, on a more realistic scale, item residual of 0.1 is associated with item parameter change of 0.189. Bank difficulty parameters and item position still remained significant in the multiple regression model, with b = 0.051 and b = 0.529, respectively. The R^2 of the model was 0.321 (F (4, 92) = 103.6, p<.001).

Table 9. Multiple Regression Results

	b	***SE***	***t***	***p***
(Intercept)	.004	.031	.125	.900
Item age	.006	.008	.800	.426
Bank parameter	.051*	.015	3.483	.001
Item position	.529*	.227	2.329	.022
Item residual	1.890*	.513	3.687	.000

*p<.05

7.4 CONCLUSION

To summarize the results, in addition to item age, bank parameters and item positions were identified as significant factors influencing item parameter change, and item residuals emerged as a useful indicator for potential drift. However, these predictors alone do not fully account for item parameter drift, underscoring the importance of ongoing item bank maintenance through re-field-testing.

Although item age was not statistically significant in the multiple regression model, it still warrants consideration given its positive correlation with parameter change. Future selection of re-field-testing items should incorporate additional predictors beyond item age, particularly item residuals, which are designed to capture potential drift. A large residual for an item with very low or high difficulty may be less concerning due to naturally larger standard errors, whereas similar residuals for mid-difficulty items could signal meaningful drift.

8. REFERENCES

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: how accurate are low stakes exams?. *Journal of Labor Research*, *42*(2), 1–60.

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest Item: calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191–212.

Baker, F. B. (1992). *Item Response Theory*. New York: Springer-Verlag.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–449.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283–296

Bock, R., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610.

Choe, E. M., Zhang, J., & Chang, H. H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika, 83*(3), 650–673.

Cui, M. (2023). *Item drift for item clusters*. Paper Presented at the 88th Annual Meeting of the Psychometric Society, Maryland, United States.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*, 265–300.

Giordano, C., Subhiyah, R., & Hess, B. (2005). *An Analysis of Item Exposure and Item Parameter Drift on a Take-Home Recertification Exam*. Paper Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 369–377.

Guo, H., Robin, F., & Dorans, N. (2017). Detecting item drift in large-scale testing. *Journal of Educational Measurement*, *54*(3), 265–284.

Han, K. T., & Guo, F. (2011). Investigating the impact of item parameter drift on CAT item calibration and selection. *Applied Psychological Measurement, 35*(6), 451–466. https://doi.org/10.1177/0146621611414219.

Kingsbury, G.G., & Wise, S. L.. Creating a K-12 Adaptive Test: Examining the Stability of Item Parameter Estimates and Measurement Scales. Retrieved from https://www.testpublishers.org/assets/documents/JATT_Vol_12_special_1_Creating_K-12_adaptive_test.pdf.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51–66.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353–368.

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147–177.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*(1), 38–47.

Robin, F., Steffen, M., & Liang, L. (2014). GRER Revised General Test (Admission Test). In D. Yan, A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 325–342). Boca Raton, FL: Chapman & Hall/CRC.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232

Swaminathan, H. & Gifford J. A. (1983). Estimation of parameters in the three parameter latent trait model. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

van der Linden W. J., Diao Q. (2011). Automated test-form generation. *Journal of Educational Measurement, 48*, 206–222.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384.

Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*(4), 373–389.

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

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

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38.

Wyse, A., & Babcock, B. (2016). How does calibration timing and seasonality affect item parameter estimates? *Educational and Psychological Measurement, 76*, 508–527.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297–311.