Journal of Computerized Adaptive Testing

Volume 12 Number 1
February 2025

Special Issue

Through-Year Assessments: Advancements and Implications for Hybrid Interim-Summative Testing

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat
ISSN: 2165-6592
DOI 10.7333/2502-1201000

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 10.7333/2502-1201001

ISSN 2165-6592

Introduction to the Special Issue

Through-Year Assessments: Advancements and Implications for Hybrid Interim-Summative Testing

Duanli Yan
Editor

The use of computerized adaptive testing (CAT) in large-scale assessments has gained widespread adoption due to its ability to optimize measurement and reduce the burden on both students and test administrators. In particular, CAT's efficiency has been central to its application in high-stakes assessments, where precision and fairness are critical (van der Linden & Glas, 2010). As the educational landscape evolves to meet the diverse needs of students and teachers, adaptive testing is increasingly being integrated into large-scale assessments to provide more timely, actionable feedback and better alignment with curriculum goals (van der Linden & Glas, 2010). These innovations have led to the development of new assessment models and new assessment systems.

One promising development in assessment is through-course summative assessment (TCSA), which combines scores from multiple tests administered throughout the academic year, instead of relying solely on an end-of-year exam. TCSA aims to provide more nuanced and continuous feedback, thus offering a more comprehensive picture of student progress. Thus, the adaptive through-year assessment system using TCSA represents a significant evolution of traditional summative assessments by providing ongoing formative feedback. The expected benefits of TCSA and through-year assessments are substantial, including finer-grained feedback (Preston & Moore, 2010), increased validity through the inclusion of performance tasks (Bennett et al., 2011), and better alignment with the curriculum (Wilson & Sloane, 2000). However, these models face numerous challenges, such as technical difficulties with score aggregation and increased complexity in designing the assessments themselves (Wise, 2011). In this context, a key innovation is the Through-Year Computerized Adaptive Test (TY-CAT), which promises to address many of these challenges by integrating the advantages of both adaptive testing and through-course models.

This special issue presents a collection of four papers that explore key challenges and innovations in the design and implementation of adaptive testing within through-course summative assessments (TCSA) and through-year assessments (TYA). These papers discuss various technical and policy considerations, offering insights into the potential of adaptive testing to enhance both summative and interim assessments.

The first paper in this issue by Gianopulos provides a literature review focusing on the benefits and challenges of through-course summative assessments (TCSA). TCSA models combine scores from assessments administered at different points throughout the school year, offering benefits such as more granular feedback, reduced measurement error, and greater alignment with curricular goals (Preston & Moore, 2010; Bennett et al., 2011). The author of this paper argues that a through-year computerized adaptive test (TY-CAT) could address many of the technical challenges inherent in traditional TCSA models. By leveraging the flexibility of adaptive testing, TY-CAT could provide more accurate and timely measurements while reducing the burden on students and teachers. This paper sets the stage for further exploration into adaptive testing's potential for continuous, high-quality assessment throughout the school year.

The second paper by Schneider et al shifts focus to the development and policy considerations involved in designing a multiple-administrations adaptive test, i.e., a through-year assessment. This paper provides a comprehensive overview of the challenges faced by stakeholders in creating a through-year assessment system that integrates interim and summative assessments into a unified framework. Using a prototype design and score reports, the authors demonstrate how through-year assessments can support both teachers in understanding student progress and states in meeting accountability requirements. The implications of this work are significant for educational policy-makers, as it suggests a pathway for developing assessments that not only meet accountability standards but also provide actionable, formative feedback that enhances instructional practice.

The third paper by Gianopulos, et al investigates a critical aspect of adaptive testing: the composition of item banks. Specifically, the authors examine the impact of different item distributions—uniform versus bell-shaped—on measurement outcomes within a hybrid interim-summative assessment model. Using simulations for Grade 4 and Grade 6 mathematics, the study explores how item bank size and distribution affect measurement precision, accuracy, and item exposure rates. This research highlights the trade-offs involved in designing adaptive tests with sufficient precision and fairness across diverse student populations. It underscores the importance of carefully managing item bank composition to ensure robust and reliable assessment outcomes while minimizing potential biases in item exposure.

Finally, the fourth paper by Lee et al. presents an innovative hybrid interim-summative adaptive assessment design that dynamically routes students to off-grade items based on their ability estimates. This design, evaluated through simulations in Grade 4 and Grade 6 mathematics, explores the feasibility of transitioning students to off-grade assessments without compromising the integrity of proficiency determinations. By incorporating both on-grade and off-grade testing, this design aims to support a more comprehensive understanding of student abilities, especially for students performing at the extremes of the achievement distribution. The paper's findings suggest that this adaptive approach could meet federal requirements while also addressing the instructional needs of diverse student populations.

The papers in this special issue underscore the potential of adaptive testing to improve educational assessments by providing more personalized, accurate, and timely measures of student performance. The significance of this research lies not only in advancing the technical aspects of

adaptive testing but also in its implications for policy, test development, and educational practice. By addressing critical issues related to test design, item pool composition, and the integration of interim and summative assessment models, this work contributes to the ongoing evolution of assessment systems that better serve the needs of students, educators, and policymakers alike.

The advancements presented in this special issue provide a clearer path forward for the development of assessments that are not only more adaptable and efficient but also better aligned with the needs of both students and educators. As we move toward more personalized and data-driven educational systems, these innovations are crucial in ensuring that assessments can provide accurate, actionable insights at multiple points throughout the academic year.

References

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Center for K–12 Assessment & Performance Management at ETS.

Gianopulos. G. (2025). A literature review of through-course summative assessment models: The case for an adaptive through-year assessment (2025). *Journal of Computerized Adaptive Testing, 12(1),* 4-34. *CrossRef*

Gianopulos. G., Lee, J., Lim, S., Niu, L., Lee, S., and Choi, S. (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. *CrossRef*

Lee. J., Lim, S., Schneider, M.C., Gianopulos, G., Niu, L., Lee., S., & Choi, S.W. (2025). The impact of item bank transition rules on student ability estimates and achievement level classifications, *Journal of Computerized Adaptive Testing, 12(1),* 88-122. *CrossRef*

Preston, J., & Moore, J. E. (2010). *An introduction to through-course assessment*. North Carolina Department of Public Instruction.

Schneider, M.C., Choi, S.W., & Lewis, D. (2025). Design considerations and reporting solutions for a multiple administrations adaptive testing system. *Journal of Computerized Adaptive Testing,12(1),* 35-53. *CrossRef*

van der Linden, W. J. and Glas, C.A,W. (2010). *Elements of adaptive testing*. Springer. New York.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13(2),* 181–208. *CrossRef*

Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.* Center for K – 12 Assessment and Performance Management at ETS.

Author Address

Duanli Yuan. Email: dyan08550@gmail.com

Citation

Yan, D. (2025). Introduction to the special issue: Through-year assessments: Advancements and implications for hybrid interim-summative testing. *Journal of Computerized Adaptive Testing, 12(1),* 1-3.

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 10.7333/2502-121004

ISSN 2165-6592

A Literature Review of Through-Course Summative Assessment Models: The Case for an Adaptive Through-Year Assessment

Garron Gianopulos
NWEA

This review describes various approaches and expected benefits of through-course summative assessment (TCSA) and many of the challenges associated with TCSA models, concluding with a case for why a through-year computerized adaptive test (TY-CAT) would solve many of the challenges. The central feature of TCSA models is that they combine scores from tests administered at different time points of the school year (U.S. Department of Education, 2010). The expected benefits of TCSAs are numerous, including finer-grained feedback due to an increase in the cumulative number of items (Preston & Moore, 2010); increased time to include and score performance tasks, which is expected to increase the content validity of summative scores (Bennett et al., 2011); increased curricular and assessment coherence (Wilson & Sloane, 2000); timely feedback (Wise, 2011); and potentially reduced measurement error (Wise, 2011). The question of whether a set of assessments could be administered throughout the school year and combined to replace a single end-of-year summative test used for accountability has been considered before. The Partnership for Assessment of Readiness for College and Career (PARCC) was considered a through-course summative design (Jerald et al., 2011). Although PARCC's proposed design created much interest initially, it brought technical challenges, and the design was changed to a more traditional summative assessment. This literature review aims to evaluate different TCSAs in the literature to learn if alternative designs, especially CAT designs, might overcome some of the technical challenges. Three blueprint designs are discussed: distributed, cumulative, and repeated comprehensive. The advantages and limitations of each blueprint and associated score aggregation methods are considered, and both technical challenges and possible solutions are reviewed. The paper concludes by considering how an interim-summative hybrid CAT addresses many of the technical challenges of TCSAs.

Keywords: *computerized adaptive testing, through-course summative assessment, through-year assessment, off-level testing, off-grade computerized adaptive tests, score aggregation*

The purpose of this literature review is to describe the advantages and limitations of various through-course summative assessment (TCSA) models with the goal of informing the design of an innovative adaptive through-year assessment system. This system is hoped to provide rich growth data and interim feedback throughout the school year while producing proficiency scores needed for state accountability at the end of the school year, replacing the end-of-year state summative assessment. The interim tests will adapt within grade to accurately assess every student against grade-level expectations, as well as above and below grade level as needed.

When TCSA was first proposed, many objections were put forth that stymied its adoption. Various TCSA models that span all modes of assessment have been proposed; however, upon studying the proposed models, it became clear that most models did not fully leverage the benefits of adaptive assessments. Consequently, many of the anticipated challenges in the literature that prevented their adoption can be mitigated by a system of adaptive assessments designed to achieve the original goals of TCSA. Therefore, before the case for an adaptive through-year assessment can be made, it is necessary to understand the original goals, designs, and challenges of TCSA. To that end, this literature review is guided by the following questions:

1. What is the definition of TCSA?
2. What are the expected benefits of TCSA?
3. What models have been proposed or discussed in the literature?
    a. What blueprint designs have been proposed by researchers?
    b. What statistical models have been proposed to combine scores from multiple interim scores into a single summative score?
4. What are anticipated challenges and potential solutions to TCSA?
5. How might adaptive tests solve the anticipated problems with TCSA?
6. What are the gaps in the literature on TCSA that need further research?

TCSA vs. a Comprehensive Balanced Assessment System (CBAS)

An important distinction to be made is between a TCSA and a CBAS. While TCSAs are most likely derived from CBASs, most CBASs are not TCSAs. TCSA is a newer concept with less appearances in the literature, and the distinction between the two systems is an important context for this review.

CBASs. A CBAS is defined in the literature as follows:

> *"Assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous. In such a system, assessments would provide a variety of evidence to support educational decision making. Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time"* (National Research Council, 2001, p. 9).

An example is the Winsight assessment system developed by Educational Testing Service (ETS) that addresses comprehensiveness, coherence, and continuity (Wylie, 2017). It is *comprehensive* in that it uses a variety of item types to measure the full range of the domain (Wylie, 2017, p. 3) and aims to address the needs of all stakeholders from the classroom to the state (Figure 1 on p. 5); it is *coherent* because it ties back to an underlying model of student learning via learning progressions (p. 3); and it is *continuous* because it includes formative, interim, and summative assessments (p. 2).

TCSAs

Even though the term "course" is used, TCSAs are applied to elementary and middle-grade content. TCSAs are defined in various ways in the literature, including the following:

> *"Academic objectives are divided into three to five units of instruction. Students take assessments on intra-year curriculum units. Unit results are aggregated to produce a summative score"* (Preston & Moore, 2010, p. 1).

The use of the term "aggregated" might give the impression that the summative score would be the simple *unweighted summation of score components* (i.e., interim scores) that measure non-overlapping content. However, a simple summation is not the only way to aggregate scores. The *Standards* define an aggregate score as "a total score formed by combining scores on the same test or across test components …[which] *may be weighted or not* [emphasis added], depending on the interpretation to be given to the aggregate score" (AERA et al., 2014, p. 215).

Even though "aggregate score" is commonly used to mean unweighted or weighted composite scores throughout the reviewed literature, the following definition is nearly identical to the former but replaces the term "aggregate" with "combined" as a different, and perhaps simpler, approach. This definition will serve as the working definition for the purpose of this literature review.

> *"Through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year"* (U.S. Department of Education, 2010, p. 18,178).

Based on this definition, a *TCSA model* is defined as a plan that answers the following two design questions.

1. How will the blueprints for each interim test be designed to ensure that the full content domain is measured by the end of the year?
2. What aggregation method will be used to combine the scores into a summative score?

"Blueprint" herein refers to a table that specifies the distribution of item score points across test events and content areas. The models reviewed in this literature review vary regarding how they answer these questions.

An example of an assessment system originally intended to be a TCSA is the Cognitively Based Assessment of, for, and as Learning (CBAL) system developed by ETS (Sabatini et al., 2011). CBAL was originally designed with "multiple events distributed across the school year... [that would] … be aggregated for accountability purposes" (Sabatini et al., 2011, p. 3). Although the CBAL system was originally designed to be a TCSA, the system was "never implemented operationally. So, details about aggregation were never worked out in practice" (J. Sabatini, personal communication, January 30, 2019).

While both Winsight and CBAL were designed to be comprehensive, coherent, and continuous and therefore could be classified as a comprehensive and balanced assessment system, Winsight does not appear to be an example of a TCSA because it does not attempt to combine scores from different points in time to produce a single summative score.

Article Selection Criteria

Articles considered for inclusion in this literature review needed to propose, discuss, or study one or more TCSA models, including a blueprint design and proposed method for combining scores. Table 1 presents the papers that satisfied these criteria. Many papers were written on the topic of TCSA circa 2010, partly in response to the U.S. Department of Education's Race to the Top Fund Assessment Program that explicitly references TCSAs (Dadey & Gong, 2017). However, very few empirical or quantitative studies have been conducted to explore the measurement challenges and advantages of TCSA. Most of these papers were concept papers and were not subjected to peer review. Dadey and Gong (2017) described the current state of the published literature on TCSA: "Developing and implementing … [TCSAs]… represent uncharted territory. Although they have been subtly promoted by the U.S. Department of Education, they have never been researched in detail nor put into practice" (p. 1). The U.S. Department of Education has promoted TCSAs most likely because TCSAs promise many advantages over traditional summative assessments, especially when considered in light of the summative assessments used in the No Child Left Behind (NCLB) era of accountability that had many unintended negative consequences along with their positives.

Table 1
Papers Included in This Literature Review

		TCSA Model		
Author(s)	Paper's Focus	Blueprint Design?	Combining Scores?	Quantitative Study?
Resnick & Berger (2010)	Proposed a TCSA model	Yes	Yes	No
Darling-Hammond & Pecheone (2010)	Proposed a TCSA model	Yes	Yes	No
Preston & Moore (2010)	Reviewed TCSA models and proposed modified TCSAs	Yes	Yes	No
Wise (2011)	Examined different TCSA blueprint schemes and score aggregation methods	Yes	Yes	Yes (simulation)
Zwick & Mislevy (2011)	Examined scaling and linking through-course	Yes	Yes	No

TCSA Model Designs

The literature includes two types of interim blueprint designs: distributed and cumulative. In distributed blueprints, the annual content is divided into discrete units designed to be administered after matching instructional units. In cumulative blueprints, each interim test measures all the content taught from the beginning of the school year up until the test event. A third alternative would be a comprehensive blueprint that repeats the same test, but none of the TCSAs reviewed in this paper described such an approach.

The score aggregation methods described in the literature can be divided into simple or complex methods:

1. Simple methods: Sum scores, maximum score, simple averages, or weighted averages.
2. Complex methods: Latent trait scale scores or expected scores based on a unidimensional item response theory (IRT) or multidimensional item response theory (MIRT) model.

Table 2 presents a matrix of seven models found in the literature based on combinations of blueprint designs and aggregation methods. In the following sections, each TCSA model is presented along with a simplified hypothetical blueprint that could be implemented with each TCSA. Each blueprint shows the distribution of items across interim tests and mathematics reporting categories. These blueprint examples are merely intended to illustrate how each TCSA might be implemented and should not be construed as the only possible designs.

Table 2
TCSA Models Based on Score Aggregation Method and Interim Blueprint Design

Summative Score Aggregation Method	Interim Blueprint Design	
	Distributed	Cumulative Distributed
Simple	1. Darling-Hammond and Pecheone's Balanced Assessment System (2010) 2. Wise's End-of-Unit Model (2011)	3. Preston & Moore's Cumulative Balanced Assessment System (2010) 4. Preston & Moore's Cumulative American Examination System (2010) 5. Wise's Continuous Learning Model (2011)
Complex	6. Resnick and Berger's American Examination System (2010)	7. Zwick & Mislevy's Cumulative Latent Trait Model (2011)

Distributed Interim Blueprints

In a distributed blueprint design, the summative blueprint is divided into mutually exclusive parts where each part is assigned to an interim time period (Preston & Moore, 2010). There are three examples of this approach in the literature:

1. Darling-Hammond and Pecheone's Balanced Assessment System (2010)
2. Wise's End-of-Unit Model (2011)
3. Resnick and Berger's American Examination System (2010)

All these models divide the total content into distinct units and assess student achievement at the end of each unit of instruction. This design is ideally suited to answer the question, "How well did a student learn recently taught content?"

Balanced Assessment System. Darling-Hammond and Pecheone's Balanced Assessment System (2010) specified curriculum-embedded performance tasks that measure complex and higher-order thinking skills in each interim test administered after one of three units of instruction.

The system gets its name from the balanced use of item types such as performance tasks, simulations, and multiple-choice items. At the end of the year, a cumulative adaptive test is administered. The performance task scores and end-of-year adaptive score are aggregated with weights to produce the summative score (Darling-Hammond & Pecheone, 2010). Table 3 illustrates a possible blueprint structure that would support this design.

Table 3
Blueprint Example: Balanced Assessment System

	Number of Points				
	Curriculum-Embedded PTs				
Reporting Category	Unit 1	Unit 2	Unit 3	End-of-Year Adaptive Test	Total
Numerical Operations	30	–	–	10	40
Algebra	–	30	–	10	40
Geometry	–	–	30	10	40
Total	30	30	30	30	120

Note. Darling-Hammond and Pecheone (2010) do not give details on how they might sample the reporting categories, so the values in the blueprint are for illustrative purposes only and do not necessarily represent the authors' intentions.

Darling-Hammond and Pecheone (2010) suggested that the total score could be a weighted combination of performance tasks and the end-of-year adaptive test score: "Student performance on the on-demand examination is intended to be combined with the embedded performance measures to contribute to a total score on the grade specific accountability measure" (p. 20). Depending on the content area and grade level, the performance tasks (PTs) would "…comprise from 20–50% of the total score" (Darling-Hammond & Pecheone, 2010, p. 20).

End-of-Unit Model. Wise's End-of-Unit Model (2011) did not specify item types but assumed that the content would be tested after each unit of instruction. This model was designed to "be a better measure of what students knew immediately after instruction in a topic or skill" (Wise, 2011, p. 19). Table 4 illustrates a possible blueprint structure that would support this design. Wise (2011) used this blueprint structure to simulate "matched scoring," meaning quarterly scores only measure what was taught in that quarter.

Table 4
Blueprint Example: End-of-Unit Model

	Number of Points by End-of-Quarter				Total Number of Points
Reporting Category	1	2	3	4	
Numerical Operations	30	–	–	–	30
Algebra 1	–	30	–	–	30
Algebra 2	–	–	30	–	30
Geometry	–	–	–	30	30
Total	30	30	30	30	120

In this model, the interim scores from each unit would be summed to arrive at a summative score used for accountability purposes. Wise (2011) conducted simulation studies that modeled different learning models, including one-time learning, one-time learning with forgetting, one-time learning with reinforcement, and learning continuously. The results of his simulation study affirmed that "…simple addition of results from each through-course assessment is appropriate" (Wise, 2011, pp. 26–27). Wise (2011) pointed out that if learning occurs after any of these interim tests, a simple summation or simple average of the scores will seriously underestimate the student's true achievement level; therefore, this design must be used for content areas in which learning is bounded to each quarter.

American Examination System. Resnick and Berger's American Examination System (2010) used a pretest and posttest design. Each posttest was a distributed accountability exam (DAE) that measured the content taught in the given unit of instruction. During each test event, the student took a posttest for the unit just taught and a pretest on the upcoming unit. This pretest/posttest design would provide a measure of academic growth through gain scores and a means for evaluating the instructional sensitivity of the test items via item gain scores. Another benefit to including pretests is that gain scores can be aggregated at the classroom or school level to produce useful data for evaluating curricula effectiveness (Resnick & Berger, 2010).

Table 5 illustrates a possible blueprint structure that would support this design. To keep the example blueprints comparable, all the blueprints in this literature review were kept at a total of 120 points. Therefore, the length of each posttest must be shorter for the American Examination System when compared to other blueprint designs to give time to the pretests. Consequently, the reliability, precision, and content coverage of the DAEs will not be as good with the inclusion of pretests unless testing time is expanded proportionally. One factor that might mitigate the problems of shorter tests is the suggestion of Resnick and Berger (2010, p. 25) to use a Bayesian latent variable model to predict future DAE scores from older DAEs, which they claimed would shorten the length of the DAEs. Nothing unique about the blueprint design would prevent this same approach from being applied to any of the TCSA models.

Table 5
Blueprint Example: American Examination System

	Number of Points						
		DAE 1		DAE 2		DAE 3	
Reporting Category	Unit 1 Pretest	Unit 1 Posttest	Unit 2 Pretest	Unit 2 Posttest	Unit 3 Pretest	Unit 3 Posttest	Total
Numerical Operations	15	15	5	5	–	–	40
Algebra	5	5	10	10	5	5	40
Geometry	–	–	5	5	15	15	40
Total	20	20	20	20	20	20	120

Note. This illustration assumes half of the items are pretest based on the statement, "If … half of each DAE's testing time were used to a pretest on the next instructional unit…" (Resnick & Berger, 2010, p. 24).

Although they did not provide an exact aggregation model, Resnick and Berger (2010) discussed the merits of a Bayesian latent variable model similar to the model used by the National Assessment of Educational Progress (NAEP). Based on the narrative, it appears they were

advocating using a weighted combination of posttest scale scores from each DAE using an IRT or MIRT model.

Advantages and Limitations of Distributed Interim Blueprints

Table 6 summarizes the advantages and limitations of distributed models. The expected benefit of the distributed blueprints is that the quality of the diagnostic feedback would be very high relative to other approaches because more testing time can be given to measure what was learned since the last interim assessment (Dadey & Gong, 2017). This approach would be more instructionally sensitive, produce equivalent scores across districts if the same pacing guide is used, and allow the summative scores to be easily summed together to arrive at a meaningful total score. However, distributed models also have several limitations to consider.

Table 6
Advantages and Limitations of Distributed Models

Advantages	Limitations
1. High-quality diagnostic feedback relative to other approaches because more testing time can be used to measure what was learned since the last interim assessment (Dadey & Gong, 2017). 2. More instructionally sensitive. 3. Can produce equivalent scores across districts if the same pacing guide is used. 4. Summative scores can be easily summed together for a meaningful total score.	1. Breadth of coverage in each interim test might be lost (Dadey & Gong, 2017). 2. The aggregated summative score might not detect knowledge that was not retained in long-term memory. 3. It does not promote retention (Preston & Moore, 2010). 4. Requires districts to use common pacing guides or common blueprints. 5. It does not support growth inferences and should only be used if academic growth is not expected to beyond the test event.

Cumulative Interim Blueprints. A criticism of the distributed blueprint approach was that it does not provide an incentive to students to retain what was learned once it has been tested. Interim blueprints that measure cumulative content address this criticism because a student's score would be lowered if they did not retain prior learning. This design is ideally suited to answer the question, "How well did a student learn and retain content?" There are four examples of cumulative design approaches in the literature:

1. Preston and Moore's Cumulative Balanced Assessment System (2010),
2. Preston and Moore's Cumulative American Examination System (2010),
3. Wise's Continuous Learning Model (2011),
4. Zwick and Mislevy's Cumulative Latent Trait Model (2011).

Cumulative Balanced Assessment System. Preston and Moore (2010) suggested a cumulative version of the Balanced Assessment System to address some of the limitations of distributed

models. This model replicates the original except that each performance task is cumulative rather than restricted to just the last unit of instruction.

Table **7** illustrates a possible blueprint structure that would support this design. Assuming that the total number of score points is fixed, one limitation of this approach is that less time will be devoted to measuring the content in the second and third instructional units because more time must be dedicated to measuring previously measured content. Moreover, it is difficult to balance content coverage in the total number of points because whatever content is taught in the first part of the school year tends to accumulate more items by the end of the year. For example, Numerical Operations includes 50 points in the total column, while Geometry has only 30 points. This might not be desirable since the proportion of items should typically match the proportion of instructional time spent on each reporting category. Preston and Moore (2010) did not provide any details on how the summative score would be produced, but they did state that methodological questions would have to be answered if this approach was used (p. 6).

Table 7
Blueprint Example: Cumulative Balanced Assessment System

Reporting Category	Number of Points				
	Curriculum-Embedded PTs			End-of-Year Adaptive Test	Total Points
	Unit 1	Unit 2	Unit 3		
Numerical Operations	30	5	5	10	50
Algebra	–	25	5	10	40
Geometry	–	–	20	10	30
Total	30	30	30	30	120

Cumulative American Examination System. Preston and Moore (2010) also suggested a cumulative version of the American Examination System. This model replicated the original American Examination System except that each DAE is cumulative rather than restricted to just the last unit of instruction.

Table 8 illustrates a possible blueprint structure that would support this design. Like the previous model, less time would be devoted to measuring the content in the second and third instructional units. It is also difficult to attain a balance of content coverage. Because testing time must be divided between posttests and pretests, fewer items are available for posttest scores that would presumably form the basis of the aggregated summative score.

Preston and Moore (2010) did not provide any recommendations for scoring the Cumulative American Examination System, but state, “This practice will raise methodological questions as to how the scores should be combined to form the student’s ‘true score’ for the year” (p. 6).

Continuous Learning Model. Wise (2011) considered multiple growth patterns, including one-time learning, one-time learning with forgetting, one-time learning with reinforcement, and learning continuously. Table 9 illustrates a possible blueprint structure that would support his Continuous Learning Model. Although Wise (2011) did not provide such details, the items within each reporting category could progress from simple to more sophisticated content across the year.

Wise (2011) compared multiple ways to aggregate scores, including simple averages, weighted averages, and maximum scores. Simple averages would place equal importance on content from

Table 8
Blueprint Example: Cumulative American Examination System

Reporting Category	Number of Points						Total Points
	DAE 1			DAE 2		DAE 3	
	Unit 1 Pretest	Unit 1 Posttest	Unit 2 Pretest	Unit 2 Posttest	Unit 3 Pretest	Unit 3 Posttest	
Numerical Operations	15	15	5	5	5	5	50
Algebra	5	5	10	10	5	5	40
Geometry	–	–	5	5	10	10	30
Total	20	20	20	20	20	20	120

each quarter, emphasizing the importance of learning each quarter's content equally well. The weighted average would emphasize content learned later in the school year, emphasizing the more sophisticated content and retention. The idea behind using a maximum score is to give credit to students for their best performance. Suppose students continuously learn through the school year and the interim test scores were scaled to maintain scale score equivalence. In that case, students are more likely to gain their highest scale score in the fourth quarter because, presumably, they have had more time to practice and master the content.

Table 9
Blueprint Example: Continuous Learning Model

Reporting Category	Number of Points by End-of-Quarter				Total Points
	1	2	3	4	
Numerical Operations	30	5	5	5	45
Algebra 1	–	25	5	5	35
Algebra 2	–	–	20	5	25
Geometry	–	–	–	15	15
Total	30	30	30	30	120

Based on the results of a simulation study under the Continuous Learning Model, Wise (2011) recommended weighted averages, where the weights were based on projection models that predicted summative scores. He reported that the weights were proportional to instructional time. According to Dadey and Gong (2017), Wise created a composite score: the first interim score had a weight of 0.10, the second a weight of 0.20, the third a weight of 0.30, and the fourth a weight of 0.40. Dadey and Gong compared different aggregation methods with highly correlated interim scores and reported no significant differences. This approach of using instructional time as a predictor of score performance is reminiscent of the Northwest Evaluation Association's (NWEA) practice of taking instructional time into account when developing MAP® Growth™ norms (Thum & Kuhfeld, 2020).

Cumulative Latent Trait Model. Like the Continuous Learning Model, Zwick and Mislevy's (2011) approach assumed that students accumulated more knowledge and skills in each content

area throughout the school year. Zwick and Mislevy (2011) recommended a latent trait model (Mislevy's Bayesian MIRT framework) to produce multiple scores, including but not limited to the aggregated summative score. They made multiple assumptions when proposing their MIRT model. Below is a subset of their assumptions most relevant to this review (Zwick & Mislevy, 2011):

1. Each interim assessment would measure a segment of the curriculum.
2. There must be domain sampling so that growth inferences can be made.
3. Schools would not be constrained to a particular curricular order (i.e., pacing guide).
4. Dichotomous and polytomous scoring is needed.
5. Many equivalent forms were needed.
6. Percentage proficient by subgroup must be reported.
7. The items need to be instructionally sensitive.

Table 10 illustrates a possible blueprint structure that would support this design. This simplified example assumes that all students receive the same set of 30 items for each TCSA and that no item appears in more than one TCSA. The 120 items represented in the table are assumed to constitute the mathematics domain. This model required the following data: a vector of item responses, *x*; a vector of curricular variables, *c*, representing the content student *i* was taught; and a vector of demographic variables, *d*. The general MIRT model expresses multiple subscores (Θ) as a function of *x*, *c*, and *d*:

$$p(\Theta|\mathbf{x}_i, \mathbf{c}_i, \mathbf{d}_i) \propto P(\mathbf{x}_i|\Theta, \mathbf{c}_i, \mathbf{d}_i)\, p(\Theta|\mathbf{c}_i, \mathbf{d}_i) = P(\mathbf{x}_i|\Theta)\, p(\Theta|\mathbf{c}_i, \mathbf{d}_i), \quad (1)$$

Table 10
Blueprint Example: Cumulative Latent Trait Model

	Number of Points												
	TCSA 1			TCSA 2			TCSA 3			TCSA 4			Total
Reporting Category	E	I	C	E	I	C	E	I	C	E	I	C	Points
Numerical Operations	10	–	–	2	5	3	2	6	2	2	6	2	40
Algebra	10	–	–	10	–	–	2	5	3	2	6	2	40
Geometry	10	–	–	10	–	–	10	–	–	2	5	3	40
Total	30	–	–	22	5	3	14	11	5	6	17	7	120

Note. E = elementary. I = intermediate. C = challenging. This table has been adapted from Zwick and Mislevy (2011).

where $p(\Theta|\mathbf{c}_i, \mathbf{d}_i)$ is the prior distribution of Θ for student *i* and $p(\Theta|\mathbf{x}_i, \mathbf{c}_i, \mathbf{d}_i)$ is the posterior distribution given the observed item responses and background variables. $P(\mathbf{x}_i|\Theta, \mathbf{c}_i, \mathbf{d}_i)$ is the likelihood function, incorporating the distribution of item responses given proficiency and the background variables. In Zwick and Mislevy's point of view, when estimating a student's individual score, *c* and *d* should be excluded from the scoring formula because all students should be held to the same standard regardless of *c* and *d*:

$$p(\Theta|\mathbf{x}_i) \propto P(\mathbf{x}_i|\Theta)\, p(\Theta). \quad (2)$$

However, when projecting a future individual score, *c* should be included because it represents exposure to the curriculum. For reporting purposes, it might be most useful to report expected scores on a released test form of items/tasks using Equation 3:

$$P(\mathbf{y}_i|\mathbf{x}_i) = \int P(\mathbf{y}_i|\Theta)\, p(\Theta|\mathbf{x}_i) d\Theta, \tag{3}$$

where *y* represents the items in the released test form. Equation 3 will project scores from different forms onto the same set of items/tasks, thereby producing a common metric.

Equation 4 was used to produce a single expected summative score (S_i^*) with weights (w_j) on each reporting category, where $\mathbf{x}_{i,obs}$ represents the subset of items in a particular TCSA, and a_j indicates if the student was administered the item (a_j = 1) or not (a_j = 0):

$$\begin{aligned} S_i^* &= E\left[S_i|\mathbf{x}_{i,obs}\right] = E\left[\sum_j w_j x_j|\mathbf{x}_{i,obs}\right] \\ &= \textstyle\sum_j a_j w_j x_{ij} + \int \sum_j (1-a_j) w_j P\left(x_j|\Theta\right) p\left(\Theta|\mathbf{x}_{i,obs}\right) d\Theta. \end{aligned} \tag{4}$$

Equation 5 can be used to predict future summative scores, assuming students had the opportunity to learn all the content represented by *c**:

$$\begin{aligned} PS_i^* &= E\left[(S_i^*|\mathbf{c}_i^*)|\boldsymbol{x}_{i,obs}, \mathbf{c}_i\right] = E\left[\left(\sum_j w_j x_j^*|\,\mathbf{c}_i^*\right)|\boldsymbol{x}_{i,obs}, \mathbf{c}_i\right] \\ &= \iint_{\theta\theta^*} \textstyle\sum_j \left(w_j P\left(x_j^*|\Theta^*, \mathbf{c}_i^*\right)\right) p(\Theta^*|\Theta, \mathbf{c}_i^*, \mathbf{c}_i) p\left(\Theta|\mathbf{x}_{i,obs}, \mathbf{c}_i\right) d\Theta^* d\Theta. \end{aligned} \tag{5}$$

Zwick and Mislevy (2011) pointed out that if the focus is on classification accuracy, the summative component of the test could focus on minimizing misclassification. This would make the test much shorter. Zwick and Mislevy (2011) assumed different pacing guides (p. 8), but in the scoring examples they assumed all students received instruction on the content prior to each TCSA. In this context, the authors excluded variable *d* (demographics) from scoring Equations 4 and 5 with the rationale that "…fairness dictates that demographic variables not be included…two individuals with the same set of item responses, but different demographic characteristics could receive a different score, which is clearly unacceptable…" (p. 13). However, their recommendation for *c* (curriculum differences) depended on the purpose: include *c* when projecting individual students' future scores but exclude *c* for individual scores (p. 13).

This leads to the question, "If it is unfair to hold different students to different standards by including *d*, then is it not also unfair to exclude *c* from individual scores if *c* is not under an individual's control?" On the contrary, it seems that including *c* would be the fairest way to score individual students because doing so would avoid penalizing students who did not have the opportunity to learn content for reasons beyond their control. Therefore, including *c* in the scoring formula would provide some statistical control that would avoid penalizing students who had less opportunity to learn the curricula, which would address the spirit of Standard 12.8 (AERA et al., 2014, p. 197).

Advantages and Limitations of Cumulative Interim Blueprints

Table 11 summarizes the pros and cons of the cumulative model. The cumulative blueprint approach addresses some of the weaknesses of the distributed blueprint design because it covers what was taught from the beginning of the school year to each interim test event. The cumulative approach also retains some of the benefits of the distributed blueprint design by striking a middle ground between breadth and depth. Depth of content coverage will be maximal at the first interim assessment, moderate at the second interim assessment, and minimal at the final interim assessment. However, a moderate degree of breadth of coverage will be attained in each interim assessment. Unlike the distributed design, the cumulative approach would be sensitive to loss of prior knowledge because prior content is repeatedly sampled in the blueprints. Because of this feature, students are given incentive to review and retain what was previously learned. Finally, the cumulative approach would most likely provide better classification accuracy than the distributed design because the last interim assessment provides information on the entire domain, making it less vulnerable to deflated or inflated scores from the Fall or Winter (assuming this plan is paired with a statistical model that combines the interim scores in such a way that gives more weight to the last interim assessment).

Table 11
Advantages and Limitations of Cumulative Models

Advantages	Limitations
1. Depth of content coverage would be maximal at the first interim assessment, moderate at the second interim assessment, and minimal at the final interim assessment. However, a moderate degree of breadth of coverage would be attained in each interim assessment. 2. Sensitive to the loss of prior knowledge because prior content is repeatedly sampled in the blueprints. Students would be given incentive to review and retain what was previously learned. 3. Most likely provides better classification accuracy than a distributed model because the last interim assessment would provide information on the entire domain, making it less vulnerable to deflated or inflated scores from the Fall or Winter.	1. Less instructionally sensitive as a school year progresses because more of the testing time must be given to the task of sampling content from prior assessments, so less time can be devoted to measuring the most recently taught content. 2. Scoring cannot be a simple summation of interim test scores because the interim test scores are not mutually exclusive. To combine the interim scores, weights would need to be applied to create a coherent and meaningful score. 3. Like the distributed models, all but one plan assumed the same pacing guides for all districts and the same blueprint design. 4. Zwick and Mislevy's model requires a special data collection effort, e.g. curricular coverage and/or opportunity to learn surveys.

A major drawback to the cumulative approach is that it will be less instructionally sensitive as the year progresses because more and more of the testing time must be given to the task of sampling content from prior assessments, so less time can be devoted to measuring the most recently taught

content (assuming test length remains the same in each interim assessment). A counterargument is that instructional sensitivity is more important and more useful in the Fall and Winter and less relevant in the Spring because little, if any, time remains for instruction following the Spring test. Another drawback to this approach is that scoring cannot be a simple summation of interim tests because the scores are not mutually exclusive parts. To combine interim scores, weights would need to be applied to create a coherent and meaningful score. Except for Zwick and Mislevy, all the researchers seemed to assume that one blueprint would work for all interim tests across all districts. However, in practice, different districts will desire different pacing guides.

Recommendations from the Literature. Wise (2011) provided many recommendations at the ETS-sponsored Through-Course Summative Assessment Symposium held in 2010 that are worth repeating here:

> *"Be very cautious in promoting or supporting uses of individual student results. Even with highly reliable tests, there will be significant measurement error in estimates of student proficiency at any one time and in measure of growth relative to some prior point of assessment. Research, likely using a test-retest design, will be needed to demonstrate that within- and between-student differences are real and not just a result of measurement error"* (Wise, 2011, p. 26).
>
> "*Methods used for aggregating results from through-course assessments to estimate end-of-year proficiency or annual growth should be based on proven models of how students learn the material that is being tested. Research…is needed to demonstrate relationships between time of instruction and student mastery of targeted knowledge and skills…mid-year results can significantly underestimate or, in some cases, overestimate end-of-year status and growth if the method for aggregation is not consistent with how students actually learn*" (Wise, 2011, p. 26).
>
> "*An end-of-unit testing model, with simple addition of results from each through-course assessment is appropriate if most or all student learning on topics covered by each assessment occurs in the period immediately preceding the assessment. Developers should also be clear whether the target is measuring maximal performance during the year or status and growth at the end of the full year of instruction*" (Wise, 2011, p. 26).
>
> "*A projection model, where results from each through-course assessment are used to predict end-of-year proficiency or growth is needed where student learning on topics covered by each assessment is continuous throughout the school year. For this approach, research will be needed to determine how to weight results from each assessment to provide the most accurate estimate of end-of-year proficiency and growth*" (Wise, 2011, pp. 26–27).
>
> "*Short-term research is needed to monitor the different ways, some possibly unintended, that through-course assessment results are used. For example, the timing of instruction or of the assessments may be altered in a way that actually detracts from learning for some or all students. Materials and guidance will be needed to promote positive uses and eliminate uses and interpretations that might have negative consequences*" (Wise, 2011, p. 27).

Zwick and Mislevy (2011) provided several recommendations to the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and

Careers (PARCC) when they were considering the use of TCSAs, as summarized below. They urged the consortia to (1) acknowledge the tradeoffs between inferential demands and procedural simplicity, (2) use the pilot and field test periods to evaluate the feasibility of the complexities of the system, and (3) standardize testing policies and procedures to ensure data quality.

Recommendation 1. Recognize the tradeoffs between inferential demands and procedural simplicity. The more demands that are made of the scaling and reporting model—that it accommodate complex items of varying instructional sensitivity, for example—the more complex the model needs to be. As demands are reduced, simpler approaches become more feasible.

Recommendation 2. Take advantage of the pilot and field test periods to evaluate psychometric approaches. For example, tests of IRT model fit can help to determine whether including complex tasks in the summative assessment scale is feasible. Pilot investigations can serve to determine if the IRT and population models can be simplified, as we note in the Possible Simplifications subsection. Pilot testing can reveal whether it is possible to relax the claims for the assessment system or add constraints to the curriculum or the assessment designs so that simpler models or approximations will suffice.

Pilot testing should include the collection of response data from students who are at different points in the curriculum and who have studied the material in different orders. This data collection would allow exploration of the dimensionality of the data with respect to the time and curricular exposure variables that must be accommodated in the TCSA paradigm. Only by examining data of this sort can we learn whether simpler IRT models can be employed. Estimation of parameters for extended response tasks, including rater effects, should be studied in pilot testing as well, since these items tend to be unstable and difficult to calibrate into existing scales. How well will they work in the anticipated system?

A data collection of this kind would also support explorations of the estimation of the posterior distribution of proficiency, $p(\Theta \mid c_i, d_i, \Gamma)$. How much data is needed for stable estimation? Are effects for c_i small enough to ignore? Again, data collection at a single occasion will not be sufficient to investigate these issues.

Finally, pilot testing should gather some longitudinal data from at least a subsample of students for purposes of studying growth modeling and combining results over occasions. Little is known about either the stability or the interpretability of results in this context.

Recommendation 3. For any assessments used to make comparisons across schools, districts, or states, recognize the importance of establishing and rigorously enforcing shared assessment policies and procedures. The units to be compared must establish policies concerning testing accommodations and exclusions for English language learners and students with disabilities, test preparation, and test security, as well as rules concerning the timing and conditions for test administration (see Zwick, 2010). Careful attention to data analyses and application of sophisticated psychometric models will be a wasted effort if these factors are not adequately controlled (Zwick & Mislevy, 2011, pp. 27–28).

Expected Advantages, Challenges, and Potential Solutions to TCSAs

The literature has pointed out many expected advantages of a TCSA compared to traditional summative tests, including the following:

1. Finer-grained feedback due to an increase in the cumulative number of items used in the calculation of summative scores (Preston & Moore, 2010).
2. Increased time to score performance tasks, which is expected to increase the content validity of summative scores since they can include more items requiring human scoring such as writing, listening, and speaking (Bennett et al., 2011).
3. Increased curricular and assessment coherence because teachers are more likely to see the connections between instruction, standards, and test items (Wilson & Sloane, 2000).
4. Timely feedback because through-year scores will be provided after each through-year test, providing teachers with the time and information they need to address students' learning needs, which is very limited with traditional summative tests (Wise, 2011).
5. Potentially reduced measurement error because of the increased number of items used for summative scores (Wise, 2011).
6. Potentially increased instructional time, assuming that interim TCSAs replace existing interim and summative tests.

The TCSA model also has several challenges, summarized in the sections below along with potential solutions.

1. Controlling for curricular exposure and opportunity to learn (OTL) can be challenging (Zwick & Mislevy, 2011; Wise, 2011). If a single blueprint is used and different districts follow different pacing guides, some students might be tested on content they did not have an opportunity to learn. This violates Standard 12.8, which stipulates that "evidence should be provided that students have had an opportunity to learn the content and skills measured by the test" (AERA et al., 2014, p. 197). The *Standards* also state, "Until such documentation is available, the test should not be used for their intended high-stakes purpose" (AERA et al., 2014, p. 189).
2. If the blueprints do not cover cumulative content, the summative score is expected to measure short-term rather than long-term retention (Nellhaus, 2010; Zwick & Mislevy, 2011). If blueprints are cumulative, the tests might take more time than users would like.
3. The peer review guidelines might impose test administration requirements that are a burden to districts (Dadey & Gong, 2017; Zwick & Mislevy, 2011).
4. Selecting the optimal score aggregation method and blueprint design is challenging because different methods might have advantages or disadvantages in different growth trajectories (Wise, 2011). Understanding how students grow differently in different content areas and ensuring that the aggregation method matches different growth trajectories might be difficult (Wise, 2011; Bennett et al., 2011).
5. Because scores from each interim test would feed into a summative score used for accountability purposes, educators might perceive the tests to be high-stakes, which might generate test anxiety and test preparation activities that reduce instructional time.
6. Given Wise's caution to use "proven models of how students learn (2011, p. 3)" to help choose a score aggregation method, considerable work should be done at the onset of test development to validate the model of student learning, which is not an easy task.

Controlling for OTL

Challenge. Typically, different school districts use different pacing guides, which means different content is covered at different points in time. If one blueprint is used for all test events throughout the year, some students will not have been taught content that will be tested, which violates Standard 12.8. (Single comprehensive blueprints can be used in traditional summative models without violating this standard because by the time students are tested, all content should have been presented and the pacing is considered irrelevant.) In addition to violating this standard, producing summative scores based on content that students did not have an opportunity to learn is fundamentally unfair. In traditional summative models, it is assumed that all students have been taught the grade or course content by the time the summative test is given at year's end. However, this assumption does not hold for each interim assessment. Therefore, the first challenge is to ensure that students have had the opportunity to learn the content being tested, or at least minimize and control the effects of not having an opportunity to learn within the aggregated summative score.

Potential solution. OTL can be controlled physically or statistically. Physical control means that the only items administered to students are items that measure content they had a high probability of being taught. This could be accomplished by developing custom interim blueprints that match the pacing guides of each district or by requesting that districts reach consensus on a single pacing guide and associated blueprint (Dadey & Gong, 2017). Different blueprints could be created for each district by collecting pacing guide information in advance and only delivering items that align to the pacing guides by adding such constraints to the constraint engine.

Alternatively, statistical control could be used by giving all students items from the same blueprint at each through-year test, collecting information from teachers concerning the opportunity their students had to learn the tested curricula, and then doing one of two things:

1. Remove items that the students did not have an opportunity to learn from the calculation of the total score,
2. Down-weight the items the students did not have an opportunity to learn.

In the statistical approach, a single comprehensive blueprint governs all interim tests and is administered to all students. Students might see items that measure content they did not learn, but the item scores are not included in the total score. Consequently, the total score only or largely reflects the content the students had an opportunity to learn. The items that the student did not have an opportunity to learn would not necessarily be wasted, for they could be combined into a subscore and used as pretest items for use in a growth model, as is promoted in Resnick and Berger's American Examination System (2010).

Another option is to down-weight the items that measure content students had no opportunity to learn to minimize their role in the aggregated summative score. Wise (2011) reported positive results when weighting interim scores proportional to the number of instructional days. Zwick and Mislevy (2011) recommended studying the effect that different curricula and instructional effects might have on aggregated summative scores to determine if the size of the effects are small enough to simply ignore. Zwick and Mislevy also discuss "MIRT models that accommodate differential change in item characteristics resulting from different ... [opportunities to learn curricula]" (p. 10).

Short-Term vs. Long-Term Retention

The challenge. Users of a TCSA summative score might interpret the scores as if the data were collected at a single point in time and therefore represent a student's achievement at the end of the year. However, if two-thirds of the data were collected from interim tests administered in the Fall and Winter, the end-of-year summative score will actually represent achievement at different points in time. This creates murkiness in the interpretation of through-year scores unless achievement does not change over time (Bennet et al., 2011). Moreover, if a Spring administration does not retest content from the first interim period(s), the score will not reflect forgotten content (Preston & Moore, 2010). The greater the gap in time between an interim test and the end-of-year summative score report, the greater the chance that the student's actual achievement level has changed.

Potential solutions. This challenge could be addressed by committing to one or the other interpretation and clearly endorsing and communicating the chosen interpretation: either attributing achievement from each interim test only to the time period it measured or explicitly designing each interim test to measure cumulative knowledge. For example, if it is intended that the summative score reflects the students' actual standing in the content standards in the Spring of the school year, the Spring interim assessment needs to have a comprehensive blueprint. If the blueprint only measures the last trimester of instruction, the score will likely overestimate or underestimate the student's actual level of knowledge of the entire domain. A comprehensive blueprint samples content from the entire school year, whether the student has an opportunity to learn the content or not. In contrast, if the summative score is not intended to represent the student's standing in the full content domain during the Spring, a more appropriate blueprint would be a distributed blueprint that divides the domain into mutually exclusive sections that are each assigned to a trimester of instruction. Using a repeated comprehensive blueprint (RCB) provides the best of both worlds in the sense that both learning and forgetting (if present) will be measured at each test. Each interim test represents the students' standing on the construct at that point in time. Aggregating scores from a RCB might or might not be necessary. The downside to this approach is that testing burden would not be greatly reduced; however, CATs that use a RCB can be designed to maximize measurement efficiency, especially if off-level testing is allowed.

Peer Review Restrictions

The challenge. Dadey and Gong (2017) observed that users of interim assessments like their high degree of flexibility and convenience. However, these features might not exist in a TCSA and be forfeited by the peer review requirements for summative assessments (U.S. Department of Education, 2015). For example, many interim tests are short, do not require a high degree of standardization, can be given within a class period, can be administered by a single teacher, and do not require a high degree of test security. In contrast, summative assessments typically take three to four hours to complete and require standardized testing conditions, a test administrator and proctor, administration training, documentation of anomalous events for test security, and special audits. These requirements make summative assessments more reliable, accurate, and valid. Although the length of a TCSA probably would not be as long as a typical end-of-year summative test, it seems reasonable to assume that the requirements of peer review would also be required of TCSA test events. Unless the peer review requirements change, summative-style test

administration protocols might place a greater level of burden on educators and support staff than expected.

Dadey and Gong (2017) provide the following warning to states considering converting interim assessments into through-year assessments:

> *"Careful and realistic consideration should be given to these questions, as well as other aspects not touched upon directly here (e.g., cost, long-term maintenance). Also, states should be cognizant of the inherent risks of repurposing interim assessments for summative purposes. Doing so runs the risk of having the interim assessments subject to the same pitfalls currently faced by large scale-summative assessments. Such pitfalls could result in two competing types of interim assessments—those mandated by the state and those educators want and use. Alternatively, interim assessments could fall out of favor altogether"* (p. 16).

Potential solutions. To address this concern in the design of the adaptive through-year assessment model, test developers should contrast their existing test administration policies with those required by peer review. Differences between the test administration requirements of the current interim or summative assessment and the planned through-year assessment system should be described to determine if the more demanding test administration policies of through-year will create burdens outweighing the perceived benefits of through-year assessments from the viewpoint of the test users. It is important that test users are educated concerning the test administration requirements of any summative assessment, including this through-year assessment system. They might expect to receive the benefits of a summative assessment but with the flexibility and convenience of a low-stakes non-summative test. This mismatch in expectations might create disappointment and frustration by test users if it is not carefully addressed early on.

Selecting the Optimal Score Aggregation Method and Blueprint Design

The challenge. Different models with varying assumptions will create different scores and inferences. Some models are cumulative in nature, testing cumulative information throughout the year, while others aim to measure only what has been learned since the last test event. Some models use simple averages to aggregate test scores, while others weight the scores to combine them into a single score. Some models project end-of-year proficiency, while others are multidimensional in nature. Researchers describe the following aggregation methods:

1. Simple summation (Wise, 2011),
2. Maximum score (Wise, 2011),
3. Simple averages (Wise, 2011; Dadey & Gong, 2017),
4. Weighted averages (Wise, 2011; Ho, 2011; Dadey & Gong, 2017)
5. Multidimensional latent trait models (Zwick & Mislevy, 2011).

Each of these aggregation methods calls for different blueprint designs. Distributed blueprints are ideal for content that is learned in just one interim period, while repeated cumulative blueprints would be ideal for content that is continually learned, practiced, and developed throughout the year. In certain content areas, some reporting categories might be time-limited while others might be continually learned throughout the school year, implying a hybrid model in which the blueprint design is either distributed or cumulative depending on the reporting category. Selecting among

the many options requires research and time. Criteria for evaluating the options should include measurement considerations and logistical and system constraints.

Possible solutions. Wise (2011) and Bennet et al. (2011) discuss various ways students are expected to learn content over time: some content is taught in a single interim period, while other skills are practiced repeatedly throughout the school year. These authors recommended that a score aggregation method and blueprint structure match the way students grow.

To address this challenge, historical assessment data could be used to model and simulate student growth at the reporting category level of the Common Core State Standards (CCSS), and monte-carlo simulation could be used to compare the precision and accuracy of various score aggregation and blueprint models. Monte-carlo simulation is an ideal method to evaluate the measurement properties of various score aggregation methods, giving researchers a way to quantify measurement precision and bias. The goal of such a simulation study is to answer the question, "What aggregation method and blueprint model produce the least amount of measurement error under each model of student learning?"

Unintended Consequences of a High-Stakes Perception

The challenge. Because scores from each interim test would feed into a summative score used for accountability purposes, educators might perceive the tests to be high-stakes, resulting in test anxiety, test preparation activities that reduce instructional time, and/or a narrowing of the curriculum (AERA et al., 2014, p. 189).

Potential solutions. The best antidote to these unintended consequences is a well-designed and balanced assessment system that is comprehensive, continuous, and coherent. To avoid narrowing the curriculum, the item banks must be *comprehensive* so that the full depth and breadth of adopted content standards are measured, which means including a variety of item types such as performance tasks and writing tasks. Large item banks should also be provided so that *if* teachers engage in periodic, even *continuous* test preparation, students will be repeatedly exposed to the full range of item types and the cognitive complexity of the content standards. Provided the items are fully aligned to the content standards, test preparation should only reinforce the content rather than narrowing it. Moreover, if students have repeated opportunities to learn content from well-aligned items and tasks, this might reduce test anxiety by increasing teachers' and learners' confidence.

To address the need for *coherence*, learning progressions can be integrated into a TCSA. Many thought leaders have pinned their hopes on learning progressions to bring much needed coherence (Resnick & Berger, 2010; Marion et al., 2018). Shepard et al. (2018) and Wilson (2018) have argued that learning progressions can improve the coherence of instruction and assessment. Marion et al. argued that learning progressions can act "as the organizing framework for connecting the various assessments and learning activities in a vertically coherent system" (2018, p. 3). Although there has been considerable optimism around learning progressions, there are also challenges with implementing and validating them.

Building Assessments on Unvalidated Learning Progressions

The challenge. Learning progressions are frequently referenced in TCSA research papers (Wise, 2011; Resnick & Berger, 2010; Zwick & Mislevy, 2011). Learning progressions are described as the "underlying model of learning" of TCSAs. There are a variety of learning

progressions and many definitions referenced in the literature (Dupree, 2011), most of which resemble the following: learning progressions "describe successively more sophisticated ways of reasoning in a content domain that follow one another as students learn" (Smith et al., 2006, p. 2). They can also describe levels of student thinking (Clements & Sarama, 2014).

Learning progressions offer many benefits, but some types of learning progressions that are curriculum-dependent might not be useful for a test intended for different school systems, states, and populations. For example, learning progressions that require educational systems to modify or change their existing pacing guides might be rejected because of the effort and resources invested in the pacing guides and associated professional development. Another challenge is that learning progressions need to be empirically validated, a timely and costly undertaking (Shavelson & Kurpius, 2012). Typically, learning progressions are developed a priori based on prior research, items are developed that align to the learning progression levels, data are collected from students, and the item difficulty patterns are examined to determine if the data empirically agree with the expected item difficulty patterns. If the patterns of empirical item difficulties agree with the predicted patterns of item difficulties, the learning progression is considered validated. However, when the empirical item difficulties contradict the expected order of the learning progression levels, which often happens for the levels near the middle of the learning progression, this problem is called "the messy middle" (Confrey et al., 2017, p. 1). Messy middles make it difficult to locate an individual student within a learning progression, undermining their utility and challenging their validity.

Assessments that depend on learning progressions have been criticized for not generalizing well to school systems that use different curricula (Y. Thum, personal communication, December 2018). They have also been criticized for failing to correctly classify students into learning progression levels (Dupree, 2011). Even the CCSS learning progressions have not been empirically validated (Pearson, 2013). Until learning progressions have been fully validated and shown to be generalizable, it might be risky to use them as the foundation for an entire assessment system, as they are likely to change during the validation process (Shavelson & Kurpius, 2012) and might not generalize across school systems.

Potential solutions. Even though learning progressions can be developed a priori and treated as theories that are empirically tested using confirmatory techniques, they can also be developed solely with empirical data using exploratory techniques. Much of the criticism leveled at learning progressions is based on research conducted with psychometric models that have strong assumptions (e.g., conditional independence, unidimensionality). However, advances in modeling techniques that require fewer assumptions might be more successful in modeling learning progressions. For example, Bayesian networks can be used to connect all the items in the item bank and link together items, content standards, and learning progressions (West et al., 2012). In this approach, directed acyclic graphs are used to define learning paths and nodes to form a network that describes existing item inter-dependencies and item difficulty patterns. Cross-validation techniques can be used to ensure that the network is reproducible and generalizable across schools, districts, and states. The network can be dynamic in the sense that as more data are collected, the network can be updated, growing as the item bank increases. Given the dynamic nature of a Bayesian network, the score reporting system must be flexibly designed to accommodate updates as more data are collected. All the paths would lead to the learning progression such as range achievement level descriptors (RALDs). In fact, RALDs can be thought of as "micro learning progressions" (P. Meyer, personal communication, January 22, 2019)

because they describe how student thinking progresses from naïve to sophisticated levels of reasoning about a content area. In this way, the network can provide instructional recommendations to teachers by identifying RALDs within a student's *zone of proximal development*, or "content which the student is ready to learn" (Dupree, 2011, p.1).

With RALDs at the center, system coherence will likely increase because "…the interpretive underpinnings used to understand where a student currently is in their learning can be based on a common set of RALDs regardless of whether the teacher uses a classroom, interim, or summative assessment" (Schneider & Nichols, 2019, p.17). RALDs are central to test development and score interpretation in a principled test design approach (Schneider & Nichols, 2019) in which "the evidence to draw conclusions is made explicit in the RALDs and items are developed specific to those evidence pieces" (Schneider & Johnson, 2019). While conventional learning progressions might contradict the order of particular pacing guides, micro learning progressions such as RALDs might be more compatible with different pacing guides and can be tested empirically throughout the test development process.

The Case for an Adaptive Through-Year System

An important consideration for the design of an adaptive through-year system is whether scores will be aggregated. An alternative to a distributed or cumulative blueprint is a *repeated comprehensive blueprint* (RCB) that repeatedly measures the domain throughout the year but requires a certain minimal coverage of on-grade content before allowing the test to adapt off grade. An adaptive test using an RCB would not require scores to be aggregated across test events. Since scores are not aggregated across time, the Fall and Winter tests would not be considered high-stakes; these tests would serve as interim tests, while the Spring test would be the single summative test, making this solution an interim-summative hybrid CAT. If this interim-summative hybrid CAT is allowed to be variable length, stopping once a certain level of precision or classification accuracy is achieved, the test can be shorter than conventional tests.

In keeping with the original goals of a TCSA, an interim-summative hybrid CAT has a dual purpose: to classify students into achievement levels based on state-specific content standards and to measure growth. Table 12 presents an interim-summative hybrid CAT blueprint design that could achieve the goals of producing growth scores and determining on-grade proficiency. This example is targeting Grade 4 but allows for Grade 3 and Grade 5 content in each administration if needed.

Table 12
Blueprint Example: Interim-Summative Hybrid CAT Targeting Grade 4

	Number of Points by Grade									
	RCB 1 (Fall)			RCB 2 (Winter)			RCB 3 (Spring)			Total
Reporting Category	3	4	5	3	4	5	3	4	5	Points
Numerical Operations	0–10	6–10	0–10	0–10	6–10	0–10	0–10	6–10	0–10	25–30
Algebra 1	0–10	6–10	0–10	0–10	6–10	0–10	0–10	6–10	0–10	25–30
Algebra 2	0–10	6–10	0–10	0–10	6–10	0–10	0–10	6–10	0–10	25–30
Geometry	0–10	6–10	0–10	0–10	6–10	0–10	0–10	6–10	0–10	25–30

This adaptive test could contain two stages focused on different inferences. Stage 1: On-grade proficiency, and Stage 2: Growth. In Stage 1, the items are constrained to the state's on-grade content standards unless the student has demonstrated mastery in a reporting category, at which time the items can be off-grade. During Stage 2, 10–20 additional items are sampled from the domain and can be on-grade or off-grade depending on the student's momentary ability estimate. According to this approach, if a student is actually on-grade, all the items administered to the student will probably be on-grade. However, if a student is actually off-grade in one or more reporting category, they will receive a mixture of on-grade and off-grade items. Business rules would have to be developed and validated to ensure that the item selection algorithm functions as intended. When the test adapts off-grade, the items presented to the student could be constrained to the items within a strand progression. The first two tests could be time-limited, but the last test could be variable length with a stopping rule based on a minimum level of measurement precision, keeping the test as short as possible.

Other variants of this model should also be considered. For example, it might be possible that a majority of the items can serve both purposes of proficiency estimation and growth simultaneously. If so, the two distinct stages might not be needed. This represents the best-case scenario for the simplest through-year design. It would be prudent to build the system flexibly enough to accommodate a two-stage test.

Many TCSA benefits can be achieved by an interim-summative hybrid CAT using a repeated comprehensive blueprint. Like other TCSAs, information from prior tests can be used in subsequent tests without the complications of aggregated scores. For example, prior scores can inform the starting point of each succeeding adaptive test, which should improve the initial item selection and user experience, if not increase test efficiency to some degree. Score aggregation is not required if the through-year test uses a repeated comprehensive blueprint, so teachers and students are given incentive to review prior learning to retain what was learned earlier in the year. Another benefit to an interim-summative hybrid CAT is that the number of tests can be reduced, provided districts replace interim tests (typically three) and the summative test with three through-year CATs (TY-CATs). Finally, interim-summative hybrid CATs have the potential to shorten the testing seat time, provided the adaptive constraints are not too numerous and rigid. All of these benefits are possible with a well-designed TY-CAT.

Many challenges of TCSAs are also mitigated by an interim-summative hybrid CAT that allows off-grade adaptivity. First, unlike TCSAs using a distributed blueprint, only the Spring test is required to produce a summative score. If a student is absent from an interim assessment test window, they will see items from the entire summative blueprint at the next administration. A valid summative score can be produced in the Spring even if a prior score is missing. It can also be produced earlier in the year and summative proficiency determinations can be pooled if a state chooses. Moreover, the interim-summative hybrid CAT can be configured to allow off-grade adaptivity in Fall and Winter but disallow off-grade adaptivity in the Spring if a state wants the Spring test to be completely on-grade.

Second, unlike the TCSAs that use different blueprints through the year (i.e., distributed and cumulative), the RCB preserves the construct and maintains score equivalence across time. This allows growth scores to be produced. Without a consistent definition of the construct, it would be difficult to measure growth. Measuring growth is vitally important because it allows students to be encouraged by their personal progress even if they are not yet proficient. Maintaining a growth mindset promotes positive achievement emotions, self-efficacy, and motivation to learn.

Third, unlike the TCSAs reviewed in this paper, the interim-summative hybrid CAT can adapt up the scale to measure students who grow and adapt down the scale for those who regress in their learning. Advanced students can progress through the entire on-grade blueprint of the test in the Fall or Winter. If states decide to pool scores, advanced students who reach proficiency early in the year could be given enrichment activities so they are continually given opportunities to learn. Even if a state decides not to pool scores for summative determinations, there is a benefit to challenging advanced students with above-grade content to promote productive struggle. If a student is in earlier stages of learning, they could be permitted to see items that cover content from a grade below. This allows students who need more opportunities for spaced practice in retrieval to have opportunities that promote learning. Table 13 summarizes the benefits of interim-summative hybrid CATs by comparing them to typical TCSAs examined in this review.

Table 13
Comparison of Through-Course Summative Assessment (TCSA) and an Interim-Summative Hybrid CAT as a Through-Year Solution

Goals	TCSA (using a distributed or cumulative blueprint with aggregated summative score)	Interim-Summative Hybrid CAT (using a repeated comprehensive blueprint with off-grade adaptivity)
1. Can it lessen total test burden?	Yes. This is done by reducing test length using a distributed blueprint, and aggregating scores across time to produce a reliable summative score.	Yes. This can be done by replacing existing interim and summative tests with TY-CAT that maximizes measurement efficiency via prior information and off-grade adaptivity. In some states, variable length computer classification tests can be used to minimize test length if growth measures are deprioritized.
2. Can information from all test events be used to inform the final summative score?	Yes. Prior scores are combined with later item scores to produce a summative score that represents the blueprint.	Yes, to a small degree. Prior scores act as the start of each succeeding adaptive test, which improves item selection and increases test efficiency.
3. Can the summative score be produced if any test score is missing?	Probably not. TCSAs require scores to be combined from all three interim tests to produce a final score that represents the blueprint. Students could be required to make up a missed test, but only if the make-up test is in close time proximity; otherwise, this might give an unfair advantage to students who	Yes or no. This depends on the chosen scoring model. Aggregated scores could be used, but are not necessary. A valid summative score can be produced in the Spring even if a prior score is missing. It can also be produced earlier in the year if a state chooses.

Goals	TCSA (using a distributed or cumulative blueprint with aggregated summative score)	Interim-Summative Hybrid CAT (using a repeated comprehensive blueprint with off-grade adaptivity)
	have more opportunity to learn the content.	
4. Can advanced students demonstrate mastery of the on-grade content before the end of the course or school year?	No. Neither a distributed nor cumulative blueprint would permit students to see items that cover content that has not yet been taught.	Yes. Advanced students can progress through the entire on-grade blueprint of the test in the Fall or Winter, even if it has not yet been taught.
5. Can lower-achieving students demonstrate mastery of below-grade content in the Fall or Winter to inform instruction?	A distributed blueprint would now allow this, but a cumulative blueprint may cover content from previous grades that has been taught but not mastered. However, this would only be feasible with a CAT.	Yes. If a student is in earlier stages of learning, they are permitted to see items that cover content from a grade below.
6. Can the test measure what content was retained by the end of the year?	Yes and no. Some TCSA models do measure retention, but the most widely described model (the distributed blueprint approach) assumes that students do not increase or decrease in learning on previously measured content from Fall or Winter.	Yes. The adaptive algorithm adapts up the scale to measure students who grow and adapts down the scale for students who decrease in achievement. This allows students who need more opportunities for spaced practice in retrieval to have opportunities that promote learning.
7. Does the test adapt below or above grade level to maximize measurement precision and instructional feedback?	No. None of the designs reviewed in the literature discuss off-grade adaptivity.	Yes. If the adaptive constraint engine determines the student is off-grade, it will search to locate the student's position on scale even if they are below or above grade.

Goals	TCSA (using a distributed or cumulative blueprint with aggregated summative score)	Interim-Summative Hybrid CAT (using a repeated comprehensive blueprint with off-grade adaptivity)
8. Is the summative score interpretable?	No. Aggregated scores mix together achievement at different points in time and might not reflect students' retained learning and achievement at year's end. It is not clear exactly what the summative score represents.	Yes. The Spring test reflects students' retained learning and achievement at year's end.
9. Is the full domain represented in the blueprints?	Both the distributed and cumulative blueprints do not use full domain sampling at each test event; therefore, the scores do not reflect the full domain.	Yes. The repeated comprehensive blueprint uses domain sampling at each test event, preserving the construct and the meaning of the score.
10. Can growth inferences be made?	The lack of construct and scale equivalence complicates, if not undermines, the measurement of growth.	Yes. The blueprint supports growth inferences because it facilitates scale equivalence across time within grade. Use of a vertical scale allows for off-grade adaptivity, improving measurement precision for off-grade students, and therefore, yielding better growth inferences.
11. Will it work across all district pacing guides?	The distributed and cumulative blueprint assumes all districts can agree to one common pacing guide, otherwise, many different blueprints would have to be designed to match all the various pacing guides within a state.	Yes. A repeated comprehensive blueprint is curriculum and pacing guide agnostic. The scores represent the students' standing on the domain at each test event.

Gaps in the Literature

While there are many unanswered questions concerning TCSA systems, this section highlights the most salient issues that need further research:

1. Very little empirical and quantitative research has been done on TCSA models.
2. All the TCSA models reviewed herein assumed that interim tests were non-adaptive. Therefore, these models might not generalize well to adaptive interim tests, so further research is needed with adaptive TCSAs.
3. All the models except Zwick and Mislevy's cumulative latent trait model assumed common pacing guides, but in practice pacing guides will vary by district, at least to some degree.
4. There are several scoring challenges in the TCSA models that need to be addressed to ensure that score imprecision and bias are adequately controlled, especially due to the differential effects of OTL.

   a. Research should be conducted to test the sensitivity of TCSA scores to different curricula and various within-year growth patterns.
   b. Even apart from the considerable technical scoring challenges of TCSAs, it is not clear that a well-designed TCSA will produce superior score inferences than a well-designed comprehensive balanced assessment system.
5. Many of the researchers emphasized the importance of selecting scoring models that matched the type of growth that takes place within each content area. Learning progressions were repeatedly referenced as being a key component to TCSAs, but little information was provided on how learning progressions could be empirically validated.

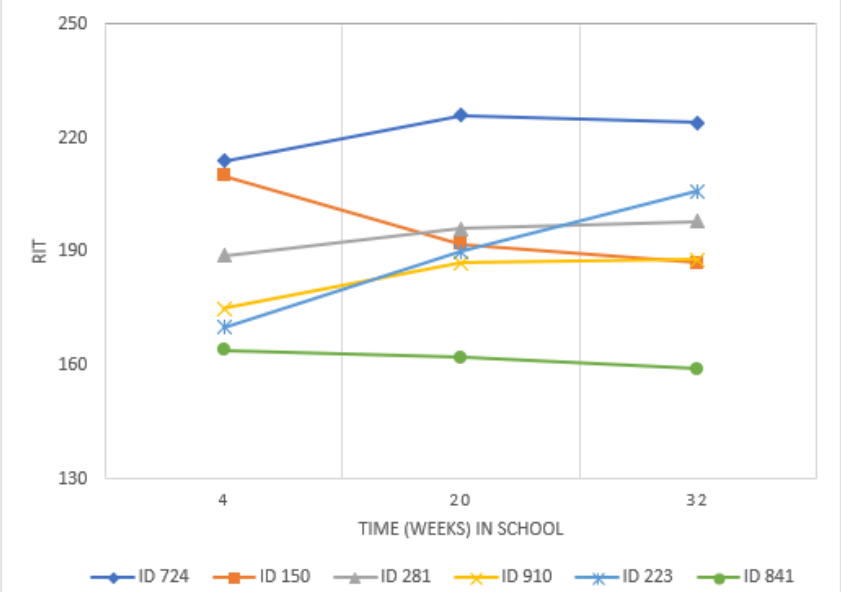
These gaps in the TCSA literature lead to the following research questions that will help guide discussions about an adaptive through-year assessment system:

1. How might the use of adaptive interim tests change the advantages and challenges of implementing a TCSA system?
2. An adaptive design would use a comprehensive interim blueprint rather than a distributed or cumulative blueprint. How might a comprehensive interim blueprint, using a repeated measures paradigm, change the TCSA approach? Does the adaptive through-year assessment system need to be a TCSA in order to achieve its intended purposes?
3. Are curricular effects such as OTL small enough to ignore? To what extent do adaptive tests minimize the negative effects of different pacing guides and OTL across districts?
4. How might a student covariate for curricular variables (*c)* be used to control for OTL in the individual summative scores?
   a. Used for predicting/adjusting IRT difficulty parameters during scoring?
   b. Used to detect differential item functioning (DIF) between no OTL and OTL during calibration?
   c. Used as a constraint variable in the adaptive algorithm?
5. What score aggregation method is best? Wise (2011) studied aggregation methods under different growth trajectories and reported that aggregation methods did not perform equally well under different growth patterns. Considering a sample of interim score patterns in Figure 1 that captures real score patterns (including patterns resembling those studied by Wise (2011)), which aggregation method produces the least amount of bias? The score patterns in Figure 1 include growth patterns that resemble typical linear growth (223, 281) but also patterns that are non-linear (724, 910) and anomalous (150, 841). These anomalous patterns are included to determine if the aggregation method will produce unbiased scores even for atypical growth patterns. How should missing interim scores be handled in the scoring methodology? What TCSA model provides the best growth measures and the best proficiency classifications?
6. Resnick and Berger (2010) suggested using prior interim scores to inform score estimation in subsequent tests. In light of this suggestion in an adaptive framework, what are the potential benefits and detriments of using prior scores as initial ability estimates (i.e., informative priors) in the adaptive engine? The algorithm needs a starting ability estimate upon which to select the first item; if that preliminary estimate is bad, the items that are selected might be less than ideal, taking longer to converge on the student's final ability estimate. If informative priors are used at the onset of the adaptive test, the adaptive algorithm will presumably converge more quickly on the student's latent trait. However, this potential benefit might backfire if the prior

ability estimate is biased. Therefore, it is prudent to ask the following: How sensitive is the constraint-based engine to a biased prior or predicted score?

7. What test lengths for the interim tests will render weighted aggregated scores that are more accurate than simply using the Spring interim assessment score as the summative test? This is important because if the Spring interim test is cumulative and provides a better measure of student achievement than a weighted score that uses Fall and Winter interim scores, the TCSA score would be inferior to simply using the last interim score. There might be a trade-off between precision and accuracy: the weighted aggregated summative score would be more stable because it is based on more information, but the last Spring interim test would be less biased because it does not contain any "outdated" information.

Figure 1
NWEA MAP Growth Interim Score Patterns



8. What role, essential or not, do learning progressions have in adaptive TCSAs? Are learning progressions generalizable enough to work across different pacing guides? How can learning progressions be empirically validated within an adaptive TCSA? How can an adaptive TCSA be developed and stabilized if it is based on learning progressions that have not yet been validated and are subject to change? How might test developers use learning progressions in the adaptive through-year assessment?
9. How does the best score aggregation method for a TCSA compare to a well-developed comprehensive balanced assessment system that does not require aggregation of scores from across the school year?

Conclusions

The purpose of this literature review was to evaluate the advantages and limitations of various TCSA models that researchers have proposed with the goal of informing the design of a new

adaptive through-year assessment system. A significant gap in the literature is a lack of research on interim adaptive tests used for TCSA models. Some of the challenges of TCSAs might be addressed via adaptive tests, but other challenges, such as the aggregation method, remain a thorny problem. Moreover, it is not clear that a weighted aggregated score combining multiple "outdated" scores will be superior to an adaptive Spring interim test. Finally, an interim-summative hybrid CAT using an RCB with off-grade adaptivity does not require score aggregation, and thus avoids many of the difficulties of TCSA. Future research, including intensive monte-carlo simulation studies and empirical, quantitative studies, should be conducted to answer the research questions raised in this paper before implementing an adaptive through-year design.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing.* AERA.

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Center for K–12 Assessment & Performance Management at ETS. *Weblink*

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach* (2nd ed.). Routledge.

Confrey, J., Maloney, A., & Gianopulos, G. (2017). Untangling the "messy middle" in learning trajectories. *Measurement: Interdisciplinary Research and Perspectives, 15*(3–4), 168–171. *CrossRef*

Dadey, N., & Gong, B. (2017, April). *Using interim assessments in place of summative assessments? Consideration of an ESSA option*. Council of Chief State School Officers. *WebLink*

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Dupree, G. (2011). *Learning progressions: A literature review*. NWEA White Paper.

Ho, A. D. (2011). *Supporting growth interpretations using through-course assessments*. Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Jerald, C. D., Doorey, N. A., & Forgione Jr., P. D. (2011). *Putting the pieces together: Summary report of the invitational research symposium on through-course summative assessments*. Center for K2 Assessment & Performance Management at ETS. *WebLink*

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2018). *A tricky balance: The challenges and opportunities of balanced systems of assessment*. National Center for the Improvement of Educational Assessment. *WebLink*

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press. *WebLink*

Nellhaus, J. (2010, January). *Race to the top assessment program: General and technical assessment discussion.* Presented at the United States Department of Education Conference on General and Technical Assessment, Washington, D.C. *WebLink*

Pearson, P. D. (2013). Research foundations of the Common Core State Standards in English Language Arts. In S. Neuman and L. Gambrell (Eds.), *Quality reading instruction in the age of Common Core State Standards* (pp. 237–262). International Reading Association. *WebLink*

Preston, J., & Moore, J. E. (2010). *An introduction to through-course assessment*. North Carolina Department of Public Instruction.

Resnick, L. B., & Berger, L. (2010). *An American examination system.* Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Sabatini, J. P., Bennett, R. E., & Deane, P. (2011). *Four years of cognitively based assessment of, for, and as learning (CBAL): Learning about through-course assessment (TCA)*. Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 13–26). Sense Publishers.

Schneider, C., & Nichols, P. (2019). *ALDs as the foundation for a coherent, integrated assessment system to support teaching and learning: The interpretive argument.* NWEA.

Schneider, M. C., & Johnson, R. L. (2019). *Using formative assessment to support student learning objectives*. Taylor and Francis. *CrossRef*

Shepard, L. A., Penuel, W. R., & Pellegrino, J. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21–34. *CrossRef*

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: a proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, *4*(1–2), 1–98. *CrossRef*

Thum, Y. M., & Kuhfeld, M. (2020). *NWEA 2020 MAP Growth achievement status and growth norms for students and schools*. NWEA Research Report. *WebLink*

U.S. Department of Education. (2010, April 9). Federal Register Volume 75, Issue 68. *WebLink*

U.S. Department of Education (2015). *U.S. Department of Education peer review of state assessment systems: Non-regulatory guidance for states*. U.S. Department of Education, Office of Elementary and Secondary Education. *WebLink*

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., Crawford, A., Choi, Y. Chapple, K. & Behrens, J. T. (2012). A Bayesian network approach to modeling learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 257–292). Sense Publishers. *WebLink*

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education,* 13(2), 181–208. *CrossRef*

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, *37*(1), 5–20. *CrossRef*

Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.* Center for K – 12 Assessment and Performance Management at ETS. *WebLink*

Wylie, E. C. (2017). *Winsight™ assessment system: Preliminary theory of action.* ETS Research Report No. RR-17-26. *WebLink*

Zwick, R. (2010). Measurement issues in state achievement comparisons (ETS Research Report No. RR-10-19). Princeton, NJ: Educational Testing Service *WebLink*
Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.* Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Acknowledgments

I wish to thank Lauress Wise and Rebecca Zwick for permission to reproduce their recommendations. I also wish to thank the NWEA leadership team for supporting this work and the following persons for providing feedback to improve the first and second version of this paper: Christy Schneider, Abby (Javarek) Andres, Yeow Meng Thum, and Sylvia Scheuring.

Author Address

Garron Gianopulos. Email: Garron@gianopulos.com

Citation

Gianopulos. G. A literature review of through-course summative assessment models: The case for an adaptive through-year assessment.
Journal of Computerized Adaptive Testing, 12(1), 4-34

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 10.7333/2502-1201035
ISSN 2165-6592

Design Considerations and Reporting Solutions for a Multiple Administrations Adaptive Testing System

M. Christina Schneider
Cambium Assessment, Inc.
Seung W. Choi
The University of Texas at Austin
Daniel Lewis
Creative Measurement Solutions LLC

This paper overviews the policy, test development, psychometric, and reporting deliberations that stakeholders engaging in the design of a multiple-administrations adaptive test, otherwise known as a through-year assessment, will have to consider. We use the development of a design prototype for simulations coupled with newly designed score reports to serve as exemplars of the work to be done by states and vendors to develop a through-year assessment with the intended purpose of merging two different assessment systems into a singular system that supports teachers in better understanding where students are in their learning and states in meeting accountability requirements.

Keywords: *computerized adaptive tests, test design, through-year assessments, score reporting*

How does the educational measurement field integrate the intended uses and purposes of interim and summative assessment systems into a single, coherent assessment system that meets the needs of state departments of education, school districts, and teachers? This is a weighty problem, oftentimes with competing goals. In this paper we overview the policy, test development,

psychometric, and reporting deliberations that stakeholders engaging in the design of such a system will have to consider. We use the development of a design prototype for simulations coupled with a newly designed mock score report as an example of the work to be done by states and vendors to develop a through-year system. Only through collaboration and innovation in the areas of (1) test design, (2) computerized adaptive test (CAT) algorithm features, and (3) score reporting features, will states and vendors merge two different assessment system purposes into a singular system that supports teachers in better understanding where students are in their learning, and states in meeting accountability requirements.

Multiple administrations adaptive tests (MAAT) in the educational measurement field heretofore have generally been referred to as interim assessments, with Curriculum Associates' *iReady*, Edmentum's *Exact Path*, Renaissance's *Star Assessments*, and NWEA's *Measures of Academic Progress* being examples. State assessments, historically, are administered only one time a year and inform state accountability. However, in 2010 the U.S. Department of Education (USDOE) described a through-course summative assessment (TCSA) in their Race to the Top applications. This has, over 15 years, slowly moved the field to considering MAAT in the context of summative purposes and uses.

Originally the USDOE (2010) encouraged the use of a TCSA.

> [A] through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year. (p. 18,178)

While stakeholders often initially like the idea of a TCSA, the research and piloting attempts over the years have shown this design has policy challenges (Gianopulos, this issue; Porter-Magee, 2011). Among the largest challenges are growth interpretations and the production of a summative score (Jerald et al., 2011). Creating a theory of action regarding how TCSA supports adapting instruction for students across the ability distributions is typically not a consideration in such a model.

The USDOE next gave flexibility in the final regulations for *The Every Student Succeeds Act* (ESSA) of 2015 (USDOE, 2016) for different assessment system designs, noting: "States have flexibility to develop new assessment designs, which may include a series of multiple statewide interim assessments during the course of the academic year that result in a single summative assessment score (sometimes described as "modular" assessments, p.3).

These interim assessments that result in a single summative score are currently referred to as through-year assessments. The USDOE (2016) clarified that innovative assessments "… may include items above or below a student's grade level so long as the State measures each student's academic proficiency based on the challenging State academic content standards for the grade in which the student is enrolled (p. 2)." USDOE's latest peer review guidance (2018) stipulated that a state can include additional content from adjacent grades in its assessments to provide additional information to parents and teachers regarding student achievement. There are, however, technical considerations for allowing students to go off grade, both above and below.

A fixed-form assessment measuring on-grade content for a state assessment will have larger measurement error at the tails of the score distribution. The students in the tails of the distribution have ability estimates that are the most imprecise, making it difficult to discern what, specifically, the student knows and can do. This situation influences the types and precision of instructional

feedback about learner profiles that are available for teachers if three fixed-form assessments are used across the year. One enhancement is to create three multistage assessments (Texas Education Agency, 2024) or three CATs using only items aligned to the target grade-level standards (Florida Department of Education, 2023). While student proficiency must be assessed using items aligned to the depth and breadth of the on-grade standards, the 2016 USDOE regulations imply that a state may be allowed to measure an outlier student more precisely by identifying where he or she is functioning, after first assessing the student's level of proficiency, so that instruction can be targeted to what the student needs next. This conclusion is derived from text noting assessments "may include items above […] a student's grade level," (USDOE, 2016 p. 2) along with the stringent requirements they have set for states to eliminate the double testing of advanced students.

The USDOE (2016) final regulations for ESSA denoted that students in eighth grade who are enrolled in Algebra may take an end-of-course test if they are taking the equivalent high school course. These students can forego taking the grade-level test of record (double testing) *only* if the state has a mechanism of providing all students "the opportunity to be prepared for and to take advanced mathematics coursework" (USDOE, 2016, p. 3). The removal of the double testing requirement is *only* allowed if the state supports the advancement and instructional supports in the same way for all students, and it sets policy that invokes subsequent changes in instruction for students across its educational system. These criteria should analogously apply to English language arts, specifically because we see states such as South Carolina enroll middle school students in end-of-course subjects in both Mathematics and English. To know which students are ready to exit grade-level standards early and enter more advanced coursework requires three criteria for a test design coupled with policy supports from the state.

1. Students can be moved off-grade with supporting evidence that the student has been measured on the breadth and depth of the on-grade-level standards and has demonstrated proficiency on grade-level standards.
2. Identified students are provided enrichment to prepare them for more advanced coursework.
3. Each year, students must again show they are meeting and exceeding the requirements for grade-level standards until they are prepared for and enrolled in high school coursework.

This policy goal translates to a design requirement for an innovative design for a through-year assessment. The prototype must allow a student to bank an advanced score on a summative blueprint and then access an above-grade-level item bank and blueprint. This would allow advanced students repeated opportunities to demonstrate and sustain advanced skills. This underpinning is related to allowing access to challenging content for all students, and it is consistent with holistic models (e.g., Assouline et al., 2009) for determining if students need acceleration. This also relates to a test design requirement that centers on first supporting a grade-level summative test score interpretation (i.e., the student should first demonstrate they are advanced in the on-grade content) and then providing guidance on where the student is functioning in the standards to support system-level and instructional-level actions through a reporting system. Because stakeholders generally want to engage in such tasks with shorter amounts of testing time, this also includes the design requirement for a CAT.

To conform with the USDOE regulations, the identification of what a student can do below-grade should also be equally evidence-based. However, the argument for students in more novice states of learning requires that these students be given access to rigorous on-grade instruction. The USDOE explicitly denoted that proficiency must be established with on-grade items, but it does

not preclude students moving to below-grade-level blueprints and item banks to support where instruction needs to begin for these students. District users of interim products frequently use test results to determine which students need to be placed into intensive Tier 3 interventions. Thus, a design requirement for an innovative design for a through-year assessment is that lower-ability students must first be measured on their present level of performance within their grade-level standards at the beginning of each test. If students in the lowest achievement level are not meaningfully accessing the grade-level standards, then these students should be routed to the adjacent below-grade-level bank and blueprint to identify where they are functioning on prerequisite standards to those in their target grade, to assist teachers in efficiently understanding how to scaffold instruction from prerequisite standards to on-grade-level standards to support student growth.

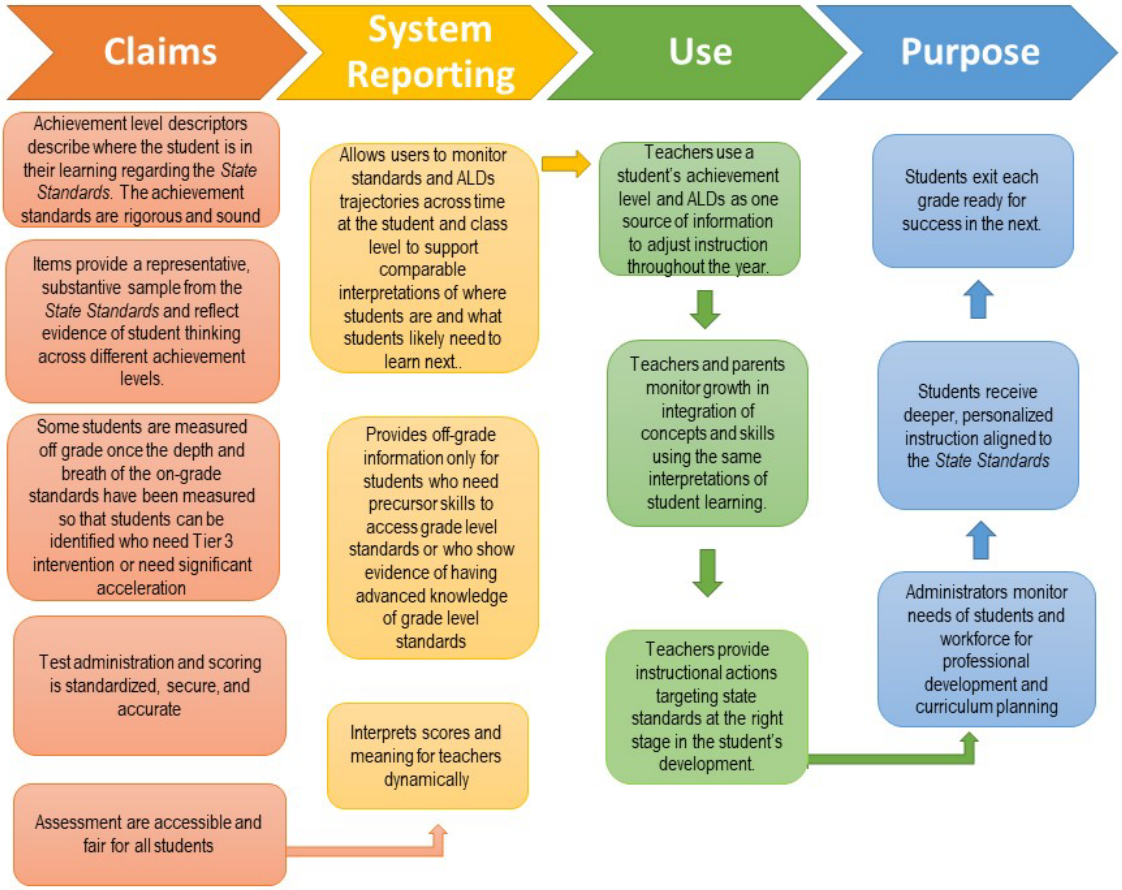
Empirically, Wei and Lin (2015) found that the measurement of students using on-grade content is accurate for most of the student population when a sufficiently large bank of on-grade content is available. However, their results show the highest scoring 10% of students and the lowest scoring 10% of students, might need off-grade content from the adjacent grades to measure their present level of functioning in the state standards. Such interpretations rely heavily on item bank depth and the psychometric qualities of the items. Test design expectations for peer review have consistently noted the need to measure the breadth and depth of the standards, which translate to a design requirement of having enough items within each achievement level bin to measure all students reliably. Because states can expect that students are growing throughout the year, it is critical to have sufficient numbers of items in the lowest and highest achievement levels to measure students on these standards at three different time points in a year with different items. This suggests test design requirements centered in Range achievement level descriptors (ALDs; Egan et al., 2012) as the score interpretation, which are then embedded into item writing, alignment, item bank analysis, standard setting, and reporting processes, along with a commitment to improving score interpretations through iteration (Huff et al., in press; Lewis & Cook, 2020; Leucht, 2020; Schneider et al., 2021). These requirements are necessary to meet the intended theory of action shown in Figure 1.

To support the intended use of monitoring growth over time there is a requirement that the test items be placed on a vertical scale with a common domain blueprint. This is an appropriate design for stakeholders who desire to support students learning at a different pace from one another; that is, they believe there is heterogeneous achievement and growth among students. Further, this design centers the goal of learning in the mastery paradigm. Guskey (2010) wrote that mastery learning (sometimes called standards-based or competency-based grading) is centered in the belief that students earn a grade (or in this context an achievement level) based on achieving mastery, and he advocated those students who need multiple opportunities to master learning targets deserve the same grade (in this case summative achievement level) as those who mastered the learning target faster. Thus, a student's summative score and proficiency is established by the end of the year for most students and can be banked earlier in the year for some students who are advanced.

When comparing the USDOE (2018) *Peer Review Guidelines* with the ESSA regulations (2016), conflicting specificity is found. Whereas the peer review requirements stipulate that the assessment "provides a *score* for the student that is based only on the student's performance on grade-level academic content standards, "(USDOE, 2018, p. 23), the ESSA regulations denote the assessment measure "each student's academic *proficiency* based on the challenging State academic content standards for the grade in which the student is enrolled" (USDOE, 2016, p. 2). The USDOE peer review guidelines also note that each student's score who is measured with off-grade content

must be as precise as the score for a student assessed only on grade-level academic content standards. This translates to a criterion that outlier scores have a conditional standard error of measurement that is similar to those students at the edges of the distribution who have tested using only on-grade content. Moreover, the state "may not include off-grade-level content in evidence addressing the critical elements" (USDOE, 2018, p. 25) for peer review, and only student performance based on grade-level academic content and achievement standards will meet accountability and reporting requirements under Title I. Thus, the final requirement for the assessment is that the prototype allow for the easy and clear extraction of data that is on grade versus off grade for accountability and that the on-grade information be sufficiently reliable for its intended purpose.

Figure 1
Theory of Action for a Principally Designed Through-Year Assessment



Translating Design Requirements to System Features for Prototype Software Development

The requirement for CAT was central to the prototype development. It was determined that creating the prototype as an extension of the shadow-test approach would be practical given this approach's ability to fully satisfy complex test blueprint requirements and the availability of an R

package implementing the shadow-test approach (Choi et al., 2022) at the time of the prototype development. The goal was to develop smaller adaptive tests to mimic summative blueprints but that strategically permitted access to off-grade items to satisfy the uses of both interim and summative test score users. Therefore, the prototype needed to be underpinned with a large, simulated item bank in sufficient numbers for each achievement level bin to measure students reliably across three assessments. The on-grade section (or module) of the assessment needed to produce a reliable maximum likelihood estimate of student ability which Davey et al. (2016) noted would require more than 15–20 items. In practice, the on-grade module of the test would need to meet a state summative assessment reliability goal of .80 or above.

In summary, there were three high-level policy goals:

1. Summative interpretations of student performance should undergird the score interpretations for each administration, such that a student who was *approaching proficient* in the winter could show comparable ability to a student who was *approaching proficient* in the spring, but who had developed that knowledge and skill faster.
2. Students could pool advanced proficiency and move on when ready.
3. Lower performing students would be allowed a clean slate at each test administration to provide them multiple opportunities to demonstrate on-grade mastery.

These goals support the intended interim use of test scores to diagnose if students are accessing or exiting on-grade content through the use of configurable routing rules to phases of the assessment, to document if students were growing. They are also intended to support the summative use of the test scores by determining the year-end achievement level of the student. In essence, rather than a multistage fixed form or a multistage assessment that dynamically routes students to a different module for the same blueprint (e.g., Luo & Wang, 2019) that increases or decreases in difficulty, the goal was to create a multistage CAT assessment in which phases shifted item bank content, if needed. The stages were described as phases during the feature development and modules in the actual software build.

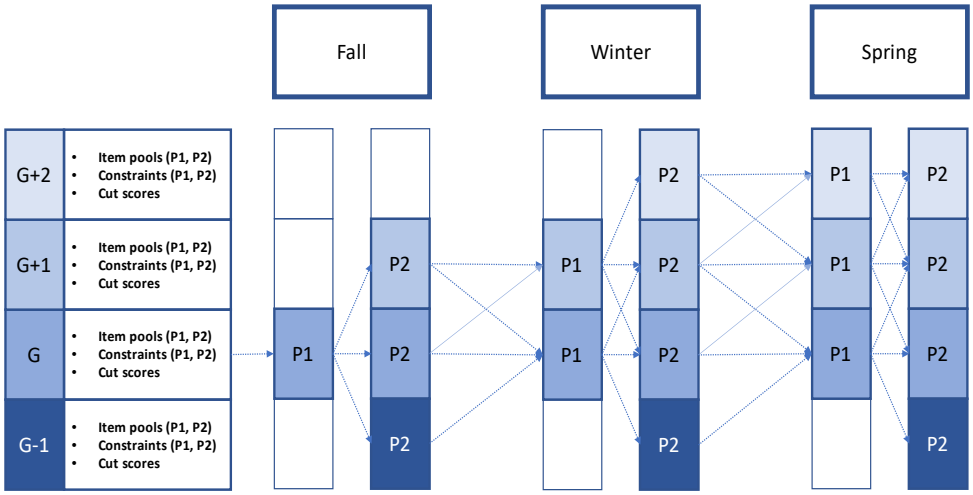
Phase Structures

The degree of off-grade adaptivity and routing rules needed to be determined. The USDOE (2016) discussed Grade 8 students taking Algebra, and Assouline et al. (2009) discussed the need to assess students in the actual above-grade content to make acceleration decisions. A system feature decision was to allow advanced students to access content in the next two higher grades sequentially if such a student met the routing criteria for their grade and the next adjacent grade. In this way, the grade-level achievement level and scale score for the student could be reported for accountability and the achievement level descriptors in the grade in which a student was functioning, along with what area of the curriculum the student needed to grow in, could be made available to help teachers prepare students for advanced coursework. For lower-ability students previously routed to the next lower grade level, a system feature decision was to always start the student in the grade-level bank. Each testing event should use the final ability estimate from the previous phase or test to initialize the subsequent phase or test. This led to the conceptualization of the desired functioning of the prototype as shown in Figure 2 (the design discussed here, while funded by NWEA, is different than the design shared in Gianopulos, this issue).

Figure 2 shows three tests, each of which has two CAT phases. In the Fall, the test comprises an on-grade-level phase and a second phase that can move off-grade, if appropriate. The shading is used to depict the grade-level bank and constraints of a phase. In the Fall, the arrows depict the

pathways to item banks and constraints that can be followed based on routing rules. The Fall has three possible pathways. One pathway is the on-grade-level phase paired with an above-grade-level phase. A second pathway is the on-grade-level phase paired with an on-grade-level phase. A third pathway is the on-grade-level phase paired with a lower-grade-level phase. The arrows between the Fall and Winter administrations show the possible pathways to the bank and constraints that begin the adaptive phase 1 of the Winter administration, which depends on the student's final ability estimate in the Fall and the routing rules. As can be seen in the figure, the number of possible pathways increase, in particular for high ability students, with each administration.

Figure 2
Routing of Phases
(From https://cran.r-project.org/web/packages/maat/vignettes/maat.html)



To achieve this design the following technical assumptions were required:

1. Within each phase, students can be routed to different grade-level blueprints and banks respectively; however, item parameters along the vertical scale can be aggregated for use in maximum likelihood scoring. This allows the final θ estimate to be comprised of item responses from phase 1 and phase 2.
2. Item banks built to Range ALDs (Egan et al., 2012) with sufficient numbers of items should allow most students in a grade to remain in the grade-level bank and show growth by moving into adjacent, higher achievement levels. This allows the majority of students to demonstrate growth in knowledge and skills while staying in the grade-level bank.
3. Item parameters on the vertical scale need to be vertically articulated such that the minimum item difficulty and the maximum item difficulty of grade G – 1 is lower than that of G. While item difficulties overlap between adjacent grades on a vertical scale, they could not do so at the tails of the distributions. This allows the use of a bank transition rule for moving off grade based on the difficulty percentile approach described below.
4. Vertical content alignment exists across domains and subdomains across grades, such that a single construct is present. This allows the final θ estimate to be comprised of item responses from phase 1 and phase 2.
5. Within and across test administrations, students should not see the same items. This allows

each test event to be comprised of new items for each student so that final θ estimates are not inflated due to item exposure.

These assumptions allow each phase to cover the content blueprint for the grade level of the phase and the corresponding item bank. This means that ultimately there is a high-level blueprint to support growth interpretations and a lower-level blueprint to support grade-level summative proficiency decisions.

Configurable Routing Rules

The design team determined that configurable routing rules should be implemented to give users different methods for determining what accessing or exiting grade-level standards meant, based on state policy. There were two approach variations stipulated for this feature: a student-centered approach and a content-centered approach. The student-centered approach was based on using confidence intervals (CI; Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000) using maximum likelihood scoring (Yang et al., 2006) extended to multiple cut scores (Thompson, 2007). The confidence interval approach was used in comparison to the cutscores for the lowest and highest achievement level in each grade. If the student's θ estimate and CI at the end of phase 1 did not overlap with the cutscore for Level 2 (for lower-performing students) or Level 4 (for advanced students) then students could be routed to an off-grade bank.

For comparison purposes, the content-centered approach was conceptualized as transitioning students when their ability was either lower than or higher than the items in the bank. This was one of the key reasons for vertically articulating the lowest and highest item parameters in the grade-level bank.

Translating System Features to Technical Specifications for Algorithms

The MAAT system is adaptive in multiple levels and built on the optimal test design framework and the shadow-test approach to CAT (van der Linden & Reese, 1998). Each test assembly within the system is performed with a clear optimality criterion and complex test specifications as a constrained combinatorial optimization problem. The test is then adapted to individual examinees through the shadow-test approach to CAT as a sequential simultaneous optimization problem. The current assessment design presumes three tests administered at specific times within a school year (e.g., Fall, Winter, and Spring) and two CAT phases within each test.

The key design features address the needs for (1) satisfying complex test specifications, (2) adapting to examinee ability within phases, (3) tracking individual examinees across test administrations, (4) controlling intra-individual item exposure, and (5) transitioning item banks between phases within a test administration and between test administrations so grade-level feedback can be provided. The system extends the shadow-test approach to CAT to assemble multiple adaptive tests optimally constructed and administered throughout the year using multiple item banks vertically scaled.

Satisfying Complex Test Specifications

A critical design requirement to support the summative use of the test scores is to maintain the same test blueprint for all students at all levels across all test administrations throughout the year. If the test design requires that a separate test blueprint be specified by test administration (or by module within each test), it should also be permissible to specify different test blueprints for the

Fall, Winter, and Spring administrations. Such a requirement might be needed when a single summative assessment is restructured into a sequence of two shorter interim tests administered throughout the year with a slightly longer summative test. For maximum flexibility, the MAAT system also supports a TCSA test design because the software allows (1) separate test blueprints specified for different test administrations (and modules within each test administration), and (2) a common test blueprint enforced for all test administrations. Given that the test assembly is performed via a mixed-integer programing solver, any complex test blueprint constraints can be satisfied while maintaining measurement optimality for individual students so long as items that meet those constraints are represented in the item bank.

Adapting to Student Ability Within Phases

While typical multistage testing presents each test module as a fixed form, the MAAT system presents each module as a fixed-length CAT, fully optimized for each student's ability using the shadow-test approach to CAT. Upon administering each item and obtaining an updated θ estimate, the system re-assembles the module to the updated θ estimate and the same test blueprint constraints. The new module contains all items previously administered within the module and new items optimizing the updated θ estimate. As a result, the new module will fully satisfy the test blueprint constraints while optimized for the updated ability estimate.

Transitioning Item Banks Within and Between Test Administrations

To further enhance the quality of measurement through the increased adaptivity, while meeting the USDOE policy guidelines, each test is designed in two phases with the provision for transitioning from one item bank (and associated test blueprint constraints) to another between the phases as determined necessary according to a prespecified transition policy. Based on an item bank and associated test blueprint constraints, each phase is a CAT assembled optimally using the shadow-test approach. At the completion of the first phase, the θ estimate from the phase determines whether the student should continue with the same item bank or be routed to an off-grade item bank in the second phase.

Figure 3 shows transition rules between phases in Test 1 and between the final θ estimate in Test 1 to the first phase of Test 2. The bank and constraints that begin the adaptive Phase 1 of Test 2 depends on the student's final θ estimate from Test 1 and the routing rules. Students can be routed to, at most, one grade level above or one grade level below between test administrations. Given G denotes the student's enrolled grade of record, any student who was previously routed to a below-grade item bank (G – 1) always starts the subsequent test on-grade, G, as shown between Test 1 and Test 2. This means, with three test administrations, the permissible item banks range from (G – 1) to (G + 2). That is, with the number of tests fixed at three per year, an advanced student can go two grades up and a more novice student in the content area can go one grade down.

As illustrated in Figure 3, transitions to different grade-level item banks can occur between phases and also between tests. In what follows, we present some details on the two approaches to implementing the above-mentioned transition rules: (1) the CI approach, and (2) the difficulty percentile approach. In both approaches, students are routed based on the performance of each phase in a test denoted as $\hat{\theta}$.

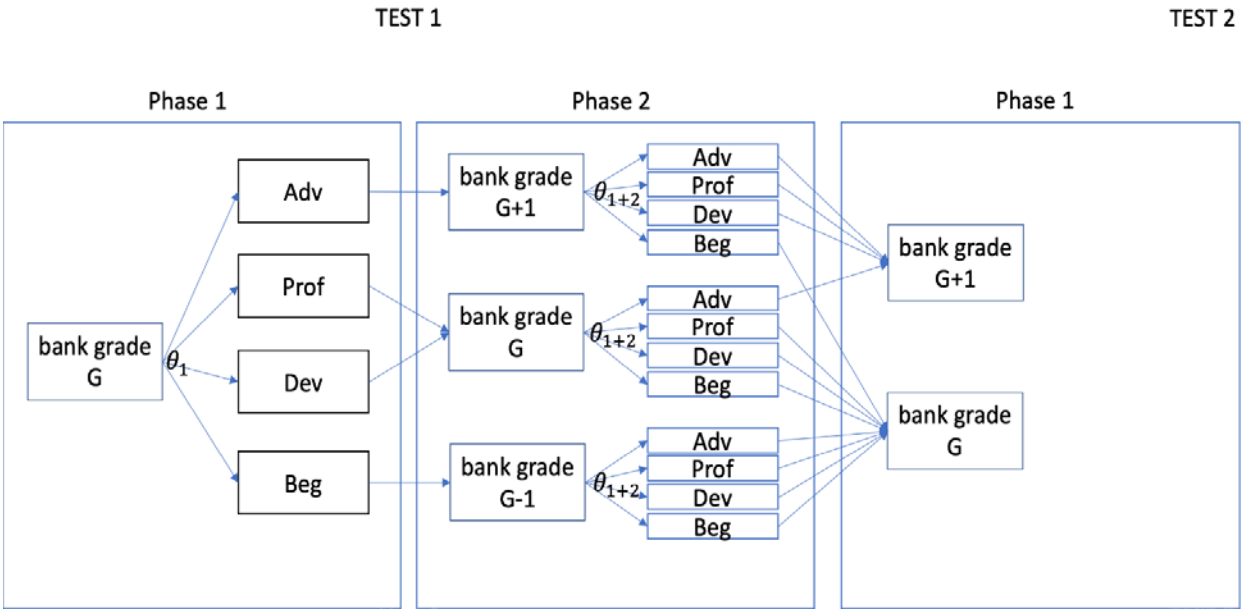
1. The CI approach computes the boundary values for each student's ability estimate:

$$\hat{\theta}_L = \hat{\theta} - z_\alpha \times SE_{\hat{\theta}} \quad (1)$$

$$\hat{\theta}_U = \hat{\theta} + z_\alpha \times SE_{\hat{\theta}} \quad (2)$$

where z_α is the normal deviate corresponding to a $(1 - \alpha)$% confidence interval, and $SE_{\hat{\theta}}$ is the standard error of measurement (SEM) associated with a point estimate of $\hat{\theta}$. Using the example

Figure 3
Transition Rules: Test 1 and Test 2



Note. θ_1 denotes an estimate based on Phase 1 items only; θ_{1+2} denotes an estimate based on a combination of Phase 1 and Phase 2 items.

with four achievement levels from Figure 3, i.e., Beginning, Developing, Proficient, and Advanced, if the lower boundary value, $\hat{\theta}_L$, falls into Advanced the student is routed to the above-grade item bank. If the upper boundary value, $\hat{\theta}_U$, falls into Beginning, the student is routed to the below-grade item bank. In all other cases, the student will remain in the same item bank.

2. If the lower boundary value, $\hat{\theta}_L$, is higher than the $(1 - \alpha)$th percentile of item difficulty values on the item response theory scale in the current bank, the student is routed to the above-grade item bank. If the upper boundary value, $\hat{\theta}_U$, is lower than the αth percentile of item difficulty values in the current bank, the student is routed to the below-grade item bank. In all other cases, the student will remain in the same item bank.

Adapting Changes in Examinee Ability Across Test Administrations

To enhance continuity and adaptivity across test administrations, each subsequent test is initialized based on the previous test performance with an opportunity to transition to a lower/higher item bank at the end of the first phase of the test. That is, in all tests Phase 1 aims to determine whether the student should be routed to an on- or off-grade bank in Phase 2. The following transition rules are illustrated in Figure 3:

1. Any student who was previously below-grade, i.e., (G – 1), always starts the next test on-

grade, G, with θ_{1+2} as the starting θ.

2. If a student is classified as above-grade after Phase 1, the next phase should be based on an above-grade bank with θ_1 as the starting θ.
3. If a student remains on-grade after Phase 1 and gets classified as above-grade after Phase 2, the next test should begin in above-grade with θ_{1+2} as the starting θ.
4. If a student transitioned into above-grade content in Phase 2, the next test should begin in the same above-grade bank with θ_{1+2} as the starting point unless θ_{1+2} fell into Beginning after Phase 2. Note that these rules are configurable in the MAAT system.
5. If a student transitioned into above-grade content in Phase 2, the next test will not move up again even if θ_{1+2} rose to Advanced.

Controlling Intra-Individual Item Exposure

The system supports options for inter-individual exposure control and intra-individual item overlap control. Exposure control is used to address test security concerns in high-stakes assessment. Overlap control, on the other hand, is used to prevent or reduce the intra-individual overlap in test content across administrations. The primary exposure control method for the shadow-test approach to CAT is the item eligibility probability method (see van der Linden & Choi, 2020). The item eligibility control method can be used to make all items previously seen by the examinee ineligible for the current administration by imposing constraints similarly as

$$\sum_{i \in S_j} x_i = 0 \,, \tag{3}$$

where s_j denotes the set of items Examinee j has seen prior to the current administration. Imposing these hard constraints can unduly limit the item bank and potentially affect the quality of measurement. To avoid infeasibility and degradation of measurement we can impose soft constraints in the form of a modification to the maximum information objective function as

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta) x_i - M \sum_{i \in S_j} x_i \,, \tag{4}$$

where M is a penalty for selecting an item from s_j, the subset of items previously administered to Examinee j. This modification to the objective function can effectively deter the selection of previously administered items unless absolutely necessary for feasibility of the model.

Although the same item eligibility constraints for inter-individual exposure control can be used to control intra-individual item exposure, the mechanism for identifying ineligible items for the intra-individual exposure control is quite different. It requires tracking the examinee records across test administrations, which might be months apart. As the number of administrations increases, the ineligible item set (s_j) can grow quickly and adversely affect the quality of measurement progressively. To prevent the ineligible item set from growing quickly, s_j might need to be defined based only on the immediately preceding test administration.

Range ALD-Based Score Reporting to Support Response to Intervention

The dynamic test design implemented in the MAAT R package (Choi et al., 2022) requires a dynamic score report structure centered in Range ALDs to encourage teachers to recognize what students need next to grow. The primary purpose of accountability assessment is to measure students' on-grade achievement and commensurately, the test design and report begin by measuring, and reporting information regarding, the students' on-grade proficiency. Figure 4 illustrates the use of principled reporting features recommended by Lewis (2019) for on-grade reporting. First, to enhance assessment literacy, we provide the most important information in question-and-answer format. That is, if tests are designed to answer questions, we can moderate the need for assessment literacy in several ways.

In particular, we explicitly state the questions and answers so that the teachers and parents do not have to make inferences. In this case, Figure 4 shows the question common for all students: *Where is this student with respect to end-of-year expectations in the grade-level curriculum?* Figure 4 illustrates how reports can answer this question dynamically, depending on each student's test performance. In this case, the answer is: *This student is currently working at Approaches Expectations.*

Another feature of the reporting structure of Figure 4 supporting the principle to enhance assessment literacy is the use of multiple reporting modalities. The question-and-answer format provides the most important information in an optimally accessible format. We also provide the students' test results analytically in two ways to support users of test results with varying degrees of analytic sophistication. We provide the information graphically, to support users capable of comprehending numerical and graphical representations of the results, and we also provide a written annotation of the results that describes in text the information that the graphics reveal.

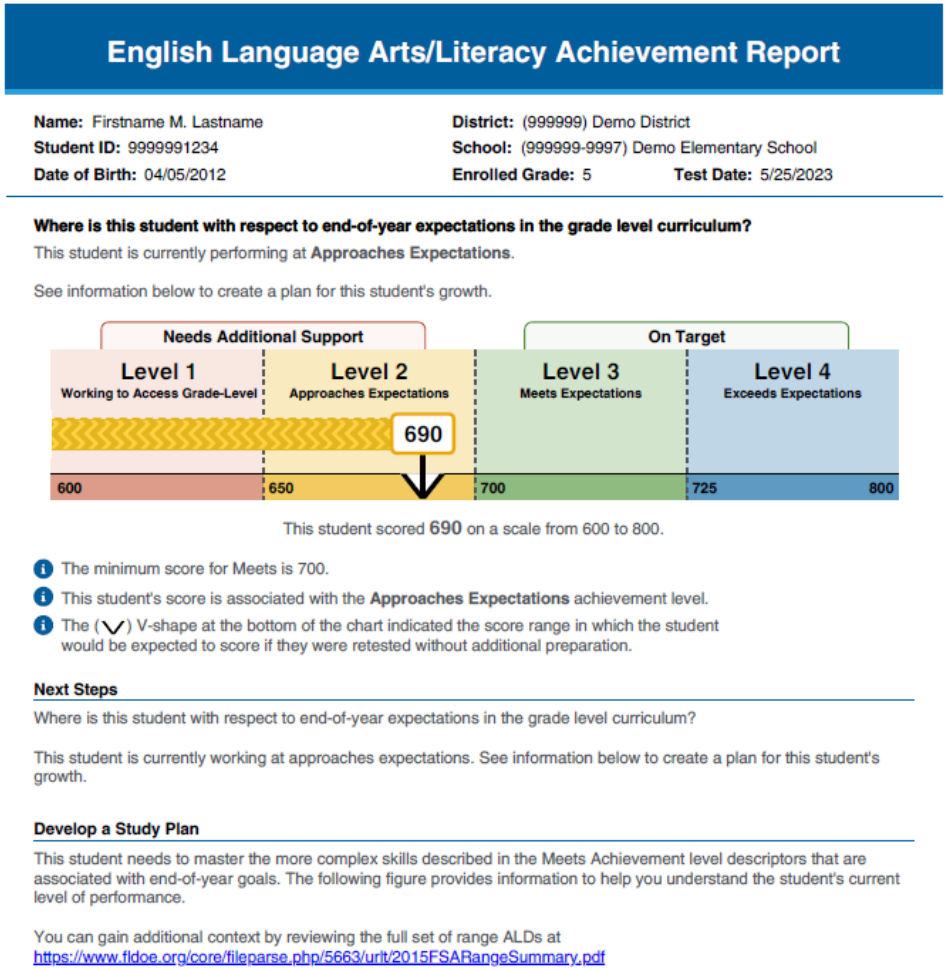
A modest, but important and often overlooked, feature illustrated in Figure 4 is the reporting and description of the meaning of the SEM in non-technical terms. It is shown as a V-shape spanning the interval of the obtained score plus and minus one SEM, supporting Lewis' (2019) primary principle—"validity first."

Following this answer to the primary question is a prompt to support another principle suggested by Lewis (2019)—enhance intelligent analytics—let teachers teach: The statement *See information below to create a plan for this student's growth* indicates that more detailed analytics follow and may be used to support student growth. The analytics that follow are also dynamic, depending on whether the student accessed fully on-grade content or on-grade and off-grade content.

Figures 5 and 6 illustrate two mechanisms for conveying more detailed test analytics for on-grade content—providing the percent correct information separately for the sets of items aligned to each performance level (Figure 5) and for the items aligned to the next higher adjacent achievement level (Figure 6) to support teachers eliciting more complex skills that the student needs next.

s

Figure 4
Sample Score Report Showing Principled Reporting Feature



English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

Where is this student with respect to end-of-year expectations in the grade level curriculum?

This student is currently performing at **Approaches Expectations**.

See information below to create a plan for this student's growth.

This student scored **690** on a scale from 600 to 800.

- The minimum score for Meets is 700.
- This student's score is associated with the **Approaches Expectations** achievement level.
- The (⌄) V-shape at the bottom of the chart indicated the score range in which the student would be expected to score if they were retested without additional preparation.

Next Steps

Where is this student with respect to end-of-year expectations in the grade level curriculum?

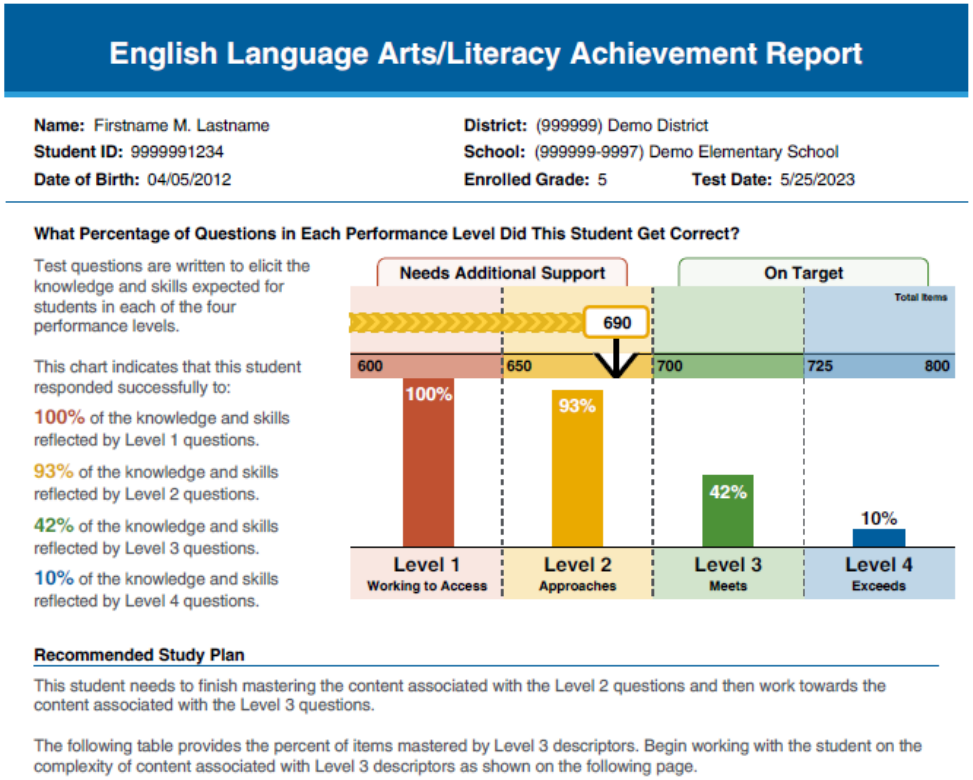
This student is currently working at approaches expectations. See information below to create a plan for this student's growth.

Develop a Study Plan

This student needs to master the more complex skills described in the Meets Achievement level descriptors that are associated with end-of-year goals. The following figure provides information to help you understand the student's current level of performance.

You can gain additional context by reviewing the full set of range ALDs at
https://www.fldoe.org/core/fileparse.php/5663/urlt/2015FSARangeSummary.pdf

Figure 5
Sample Score Report Showing Principled Reporting by Achievement Level



English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

What Percentage of Questions in Each Performance Level Did This Student Get Correct?

Test questions are written to elicit the knowledge and skills expected for students in each of the four performance levels.

This chart indicates that this student responded successfully to:

100% of the knowledge and skills reflected by Level 1 questions.

93% of the knowledge and skills reflected by Level 2 questions.

42% of the knowledge and skills reflected by Level 3 questions.

10% of the knowledge and skills reflected by Level 4 questions.

Recommended Study Plan

This student needs to finish mastering the content associated with the Level 2 questions and then work towards the content associated with the Level 3 questions.

The following table provides the percent of items mastered by Level 3 descriptors. Begin working with the student on the complexity of content associated with Level 3 descriptors as shown on the following page.

Figure 6
Sample Score Report Showing Principled Reporting of Percent of Items Aligned to Level 3 ALDs the Student Performing in Level 2 Answered Correctly

English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

How Did Your Student Perform in the Different Areas of the Test?

Reading Literature

Standard	Level 3 Descriptors	Percent Correct by Descriptor
5RL1.1	quotes accurately from a text when explaining what the text says explicitly and when drawing inferences from the text	43%
5RL1.2	determines a theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarizes the text	50%
5RL1.3	compares and contrasts two or more characters, settings, or events in a story or drama, drawing on specific details in the text (e.g., how characters interact)	50%
5RL2.4	determines the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes	42%
5RL3.4	determines or clarifies the meaning of unknown or multiple-meaning words or phrases by using context (e.g., cause/effect relationships and comparisons in a text) as a clue to the meaning of a word or phrase; uses common, grade-appropriate Greek and Latin affixes and roots as clues to the meaning of a word	40%
5RL3.5	determines understanding of figurative language, word relationships, and nuances in word meanings by interpreting figurative language, including similes and metaphors, in context; recognizes and explains the meaning of common idioms, adages, and proverbs; uses the relationship between particular words (e.g., synonyms, antonyms, homographs) to better understand each of the words	45%
5RL2.5	explains how a series of chapters, scenes, or stanzas fit together to provide the overall structure of a particular story, drama, or poem	18%
5RL2.6	describes how a narrator's or speaker's point of view influences how events are described	43%
5RL3.7	analyzes how visual and multimedia elements contribute to the meaning, tone, or beauty of a text (e.g., graphic novel, multimedia presentation of fiction, folktale, myth, poem) Also Assesses SL1.2:summarizes a written text read aloud or information presented in diverse media and formats, including visually, quantitatively, and orally	20%
5RL3.9	compares and contrasts stories in the same genre (e.g., mysteries and adventure stories) on their approaches to similar themes and topics	31%

Figure 7
Sample Score Report Showing Principled Reporting Features for a Student Currently Functioning in Level 1

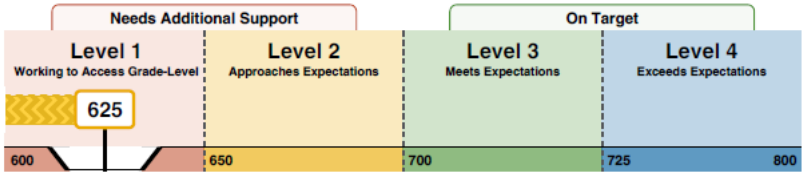
English Language Arts/Literacy Achievement Report

Name: Firstname W. Lastname
Student ID: 9999995678
Date of Birth: 04/06/2012
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

Where is this student with respect to end-of-year expectations in the grade level curriculum?

This student is currently performing at **Working to Access Grade-Level**.

See information below to create a plan for this student's growth.



This student scored **625** on a scale from 600 to 800.

- The minimum score for Meets is 700.
- This student's score is associated with the **Working to Access Grade-Level** achievement level.
- The (⋁) V-shape at the bottom of the chart indicated the score range in which the student would be expected to score if they were retested without additional preparation.

Next Steps

This assessment measures on-grade curriculum and, if the student's performance indicates, can also measure below- or above-grade curriculum. This student's achievement on the assessment indicates the student was not sufficiently accessing on-grade content. This student would likely benefit from interventions that support the student moving into the on-grade curriculum.

See the next section to develop a study plan to support student growth.

Develop a Study Plan

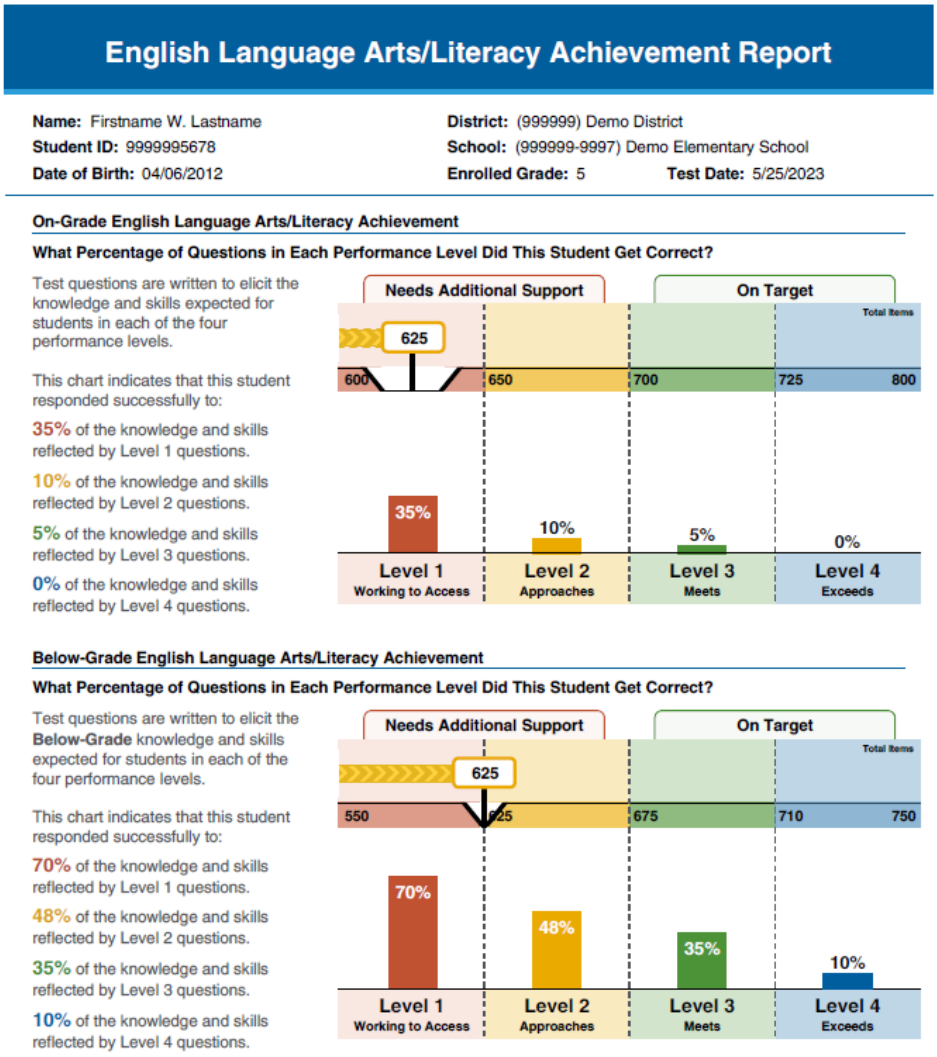
This student needs intensive support in order to master some of the skills associated with below-grade and on-grade end-of-year goals. The following two figures provide information to help you understand the student's current level of performance.

You can gain additional context by reviewing the full set of range ALDs at https://www.fldoe.org/core/fileparse.php/5663/urlt/2015FSARangeSummary.pdf

When students access on- and below-grade content, the graphic for the on-grade information is again presented; however, the next steps explicitly request the teacher provide more support to the students, as shown in Figure 7.

Depicted in Figure 8 is the breakdown of the student's performance for items by achievement level for both on-grade and below-grade content, to remind the teacher that the student needs support in precursor standards in addition to on-grade standards. This report shows teachers that a student in Level 1 in grade-level standards is likely to also function in or near Level 1 in the lower adjacent grade. Figure 9 shows a policy decision in reporting. Because the goal is to move the student into the on-grade content, we chose to report the student's performance on the Level 2 descriptors in their grade of record to encourage teachers to think about having the student move into more complex content while ensuring precursor content from the lower adjacent level is addressed. The Range ALDs are intended to show the teacher they should have the student use explicit evidence found in texts and do something with it, such as write an explanation in order to

Figure 8
Sample Score Report Showing Principled Reporting of Percent of Items by Achievement Level Both On-and Off-Grade



grow. This should help the teacher realize that the student should not just retrieve or locate explicit details. This is in contrast to the Level 3 descriptors for the student currently functioning in Level 2 who needs to grow in the Level 3 skills that are moving into inferencing. However, if these reports were produced for a student who was routed above grade level, we would expect the report to show where the student is functioning in relation to the above-grade cutscores and to show the percent of items aligned to the descriptors in the next higher achievement level in the adjacent grade to support acceleration.

Figure 9
Sample Score Report Showing Principled Reporting of Percent of Items Aligned to Level 2 ALDs the Student Performing in Level 1 Answered Correctly

English Language Arts/Literacy Achievement Report

Name: Firstname W. Lastname
Student ID: 9999995678
Date of Birth: 04/06/2012
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

How Did Your Student Perform in the Different Areas of the Test?

Reading Literature

Standard	Level 2 Descriptors	Percent Correct by Descriptor
5RL1.1	quotes accurately to support ideas stated explicitly	13%
5RL1.2	determines an explicitly stated theme from key details of a story, drama, or poem; determines the key details that should be included in a summary	20%
5RL1.3	compares and contrasts two characters, settings, or events in a story or drama, drawing on explicitly stated details in the text	21%
5RL2.4	determines the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes, through explicitly stated details	12%
5RL3.4	determines or clarifies the meaning of unknown or multiple-meaning words or phrases by using explicit context as a clue to the meaning of a word or phrase; determines the meaning of a word when given the meaning of a Greek or Latin affix or root	10%
5RL3.5	determines understanding of figurative language and word relationships in word meanings by recognizing basic figurative language, including similes and metaphors, in context; recognizes common idioms, adages, and proverbs; recognizes the relationship between particular words (e.g., synonyms, antonyms, homographs) to better understand each of the words	15%
5RL2.5	identifies the overall structure of a particular story, drama, or poem	18%
5RL2.6	states how a narrator's or speaker's point of view affects how major events are described	13%
5RL3.7	describes how visual and multimedia elements contribute to the meaning of a text Also Assesses SL1.2: determines the key details of a written text read aloud or information presented in diverse media and formats, including visually, quantitatively, and orally	16%
5RL3.9	compares and contrasts stories in the same genre (e.g., mysteries and adventure stories) on their approaches to similar stated topics	5%

Conclusions

The algorithm and business rule development process described in this paper capitalizes on the USDOE (2016) guidance that innovative assessments are permitted to include off grade items as long as the state determines if the student is proficient with on-grade items reliably. USDOE's latest peer review guidance (2018) stipulated that a state may include additional content from adjacent grades in its assessments to provide additional information to parents and teachers regar-

ding student achievement. The algorithm and business rules described here account for the technical considerations for allowing students to go off grade, both above and below, for only the specific students who are outliers. The algorithms are intended as a tool to be coupled with a test design centered in Range ALDs as the score interpretations that are embedded into item writing, alignment, item bank analysis, standard setting, and validation (Huff et al., in press; Lewis and Cook, 2020; Luecht, 2020; Schneider et al., 2021) to support improved reporting information.

Without innovation in how we (1) design and develop assessments, (2) implement CAT algorithms and business rules to control which students go off grade and when, and (3) dynamically report scores, through-year assessment systems will not meet their potential. To meet the intended uses and purposes of interim and summative assessment systems and solve the problems of bridging these two systems into a single, coherent assessment system requires that the information derived from such assessments is viewed as useful and worthy of educators' and students' time. This is critically important as educational systems work to support students whose education was disrupted by the pandemic. Helping teachers visualize that students respond successfully more often to items in lower achievement levels than to items in higher achievement levels might assist them in making connections that students need more rigor within the standards in order to grow.

References

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

Assouline, S., Colangelo, N., Lupkowski-Shoplik, A., Forstadt, L., & Lipscomb, J. (2009). *Iowa Acceleration Scale manual: A guide for whole-grade acceleration K-8* (3rd Ed.). Scottsdale, AZ: Great Potential Press.

Barnard, J. J. (2015). Implementing a CAT: The AMC experience. *Journal of Computerized Adaptive Testing*, *3*, 1–12. *CrossRef*

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Princeton, NJ: Educational Testing Service. *WebLink*

Choi, S. W., Lim, S., & van der Linden, W. J. (2022). TestDesign: An optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, *49*, 191–229. *CrossRef*

Choi, S. W., Lim, S., Niu, L., Lee, S., Schneider, C. M., Lee, J., & Gianopulos, G. J. (2022). MAAT: An R package for multiple administrations adaptive testing. *Applied Psychological Measurement*, *46*(1),73–74. *CrossRef*

Choi, S. W., Lim, S., Niu, L., Lee, S. (2022). *MAAT: Multiple administrations adaptive testing. R package* (Version 1.0.2.9000) [Computer software]. *CrossRef*

Davey, T., Pitoniak, M. J., & Slater, S. C. (2016). Designing computerized adaptive tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 483–500).

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

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

Florida Department of Education. (2023). Test design summary and blueprint: FAST ELA reading and B.E.S.T writing. *WebLink*

Jerald, C. D., Doorey, N. A., & Forgione, P. D., Jr. (2011). *Putting the pieces together: Summary report of the invitational research symposium on through-course summative assessments*. Princeton, NJ: Educational Testing Service. *WebLink*

Gianopulos, G. (this issue—2025). A literature review of through-course summative assessment models: The case for an adaptive through-year assessment. *Journal of Computerized Adaptive Testing,12 (1), 4-34. CrossRef*

Guskey. T. (2010). Lessons of mastery learning. *Educational Leadership*, *68*(2), 52-57. *WebLink*

Huff, K., Nichols, P., & Schneider, M. C. (in press). Designing and developing educational assessments. In L. Cook & M. J. Pitoniak (Eds.), *Educational measurement: 5th edition*. NCME.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). Academic Press.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Lewis, D. (2019, February). A principled approach to score reporting. Invited presentation to the CCSSO winter meeting of the Technical Issues in Large Scale Assessment (TILSA) SCASS, Baltimore, MD.

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, *39*(1), 8–21. *CrossRef*

Luecht, R. M. (2020). Generating performance-level descriptors under a principled assessment design paradigm: An example for assessments under the next-generation science standards. *Educational Measurement: Issues and Practice, 39*(4), 105–115. *CrossRef*

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing*, *19*(3), 227–247, *CrossRef*

Porter-Magee, K. (2011). *PARCC eliminates through-course assessments.* Washington, D.C.: Thomas Fordham Institute. *WebLink*

Schneider, M. C., Chen, J., & Nichols, P. (2021). Using principled assessment design and item difficulty modeling to connect hybrid adaptive instructional and assessment systems: Proof of concept. In Sottilare, R. A., & Schwarz, J. (Eds.). *Adaptive Instructional Systems. Adaptation Strategies and Methods. HCII 2021. Lecture Notes in Computer Science*, (12793). Springer. *CrossRef* .

Schneider, M. C., Agrimson, J., & Veazey, M., (2021). Examining alignment of mathematics test score interpretations on a computer adaptive assessment. *Educational Measurement Issues and Practices,41*(2), 12-24. *CrossRef*

Texas Education Agency. (2024, August 1) Texas Through-Year Assessment Pilot (TTAP) Year 1 Pilot Report. *WebLink*

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, and Evaluation*, *12*(1). *WebLink*

U.S. Department of Education (USDOE). (2010, April 9). Race to the Top Fund Assessment Program; notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register, 75*(68), p. 18,178.

U.S. Department of Education (USDOE). (2016, November 29). *Federal Register*, *81*(229). 34 CFR 200 34 CFR 299. *WebLink*

U.S. Department of Education (USDOE). (2018, September 24). *A state's guide to the U.S. Department of Education's peer review process.* *WebLink*
van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270. *CrossRef*
van der Linden, W. J., & Choi, S. W. (2020). Improving item-exposure control in adaptive testing. *Journal of Educational Measurement*, 57, 405-422. *CrossRef*
Wei, H., & Lin, J. (2015). Using out-of-level items in computerized adaptive testing. *International Journal of Testing, 15*, 50–70. *CrossRef*
Wise, L. L. (2011). Picking up the pieces: Aggregating results from through-course assessments. Center for K–12 Assessment & Performance Management at ETS. *WebLink*
Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on Multiple-category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66*, 545-564. *CrossRef*
Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.* Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Acknowledgements and Assistance

The prototype development described in this paper was funded by NWEA and is open-source. Thank you to Eli Mintzer at Cambium Assessment for his support in taking an early mock-up draft report and extending the visualizations.

Author's Address

M. Christina Schneider *Email*: Christina.Schneider@cambiumassessment.com

Citation

Schneider, M. C., Choi, S.W., & Lewis, D. (2025). Design considerations and reporting solutions for a multiple administrations adaptive testing system. *Journal of Computerized Adaptive Testing,12(1)*, 35-53.

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.

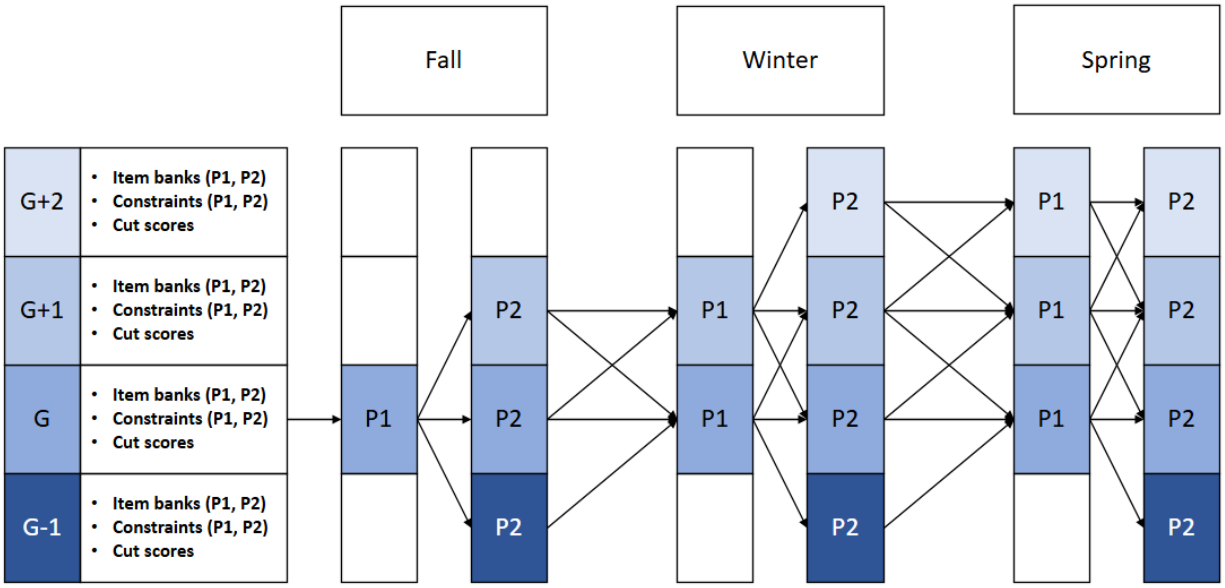
CAT Design

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

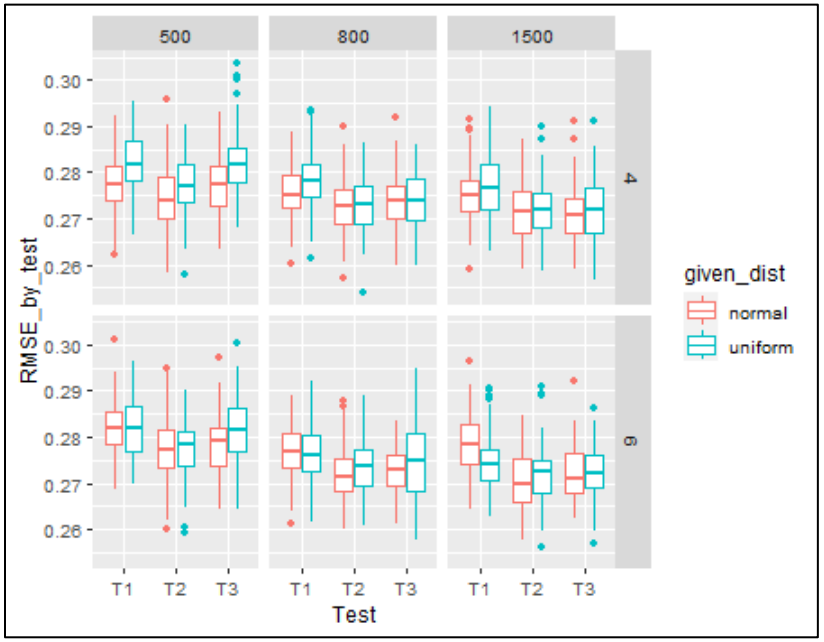


Figure 3
Bias Across Tests and Conditions

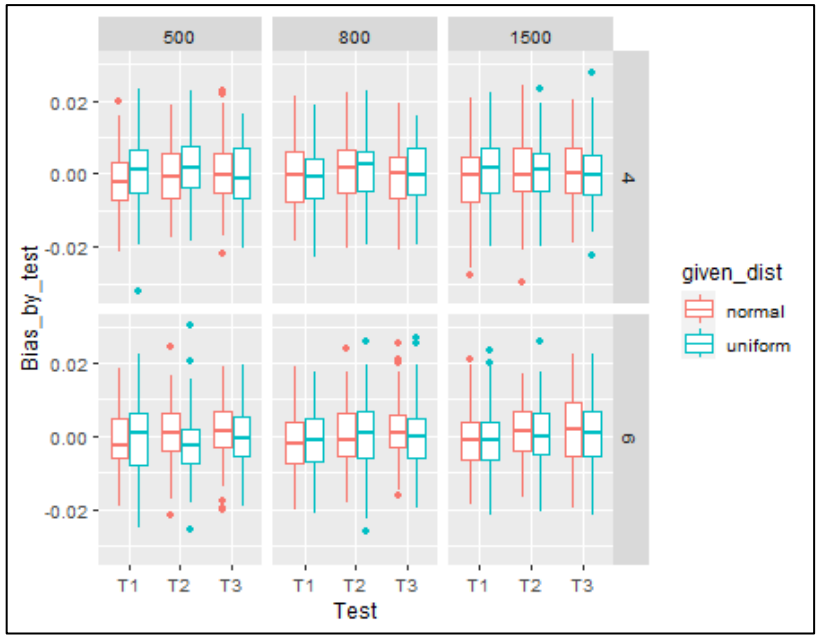


Figure 4
Ratio Index Across Tests and Conditions

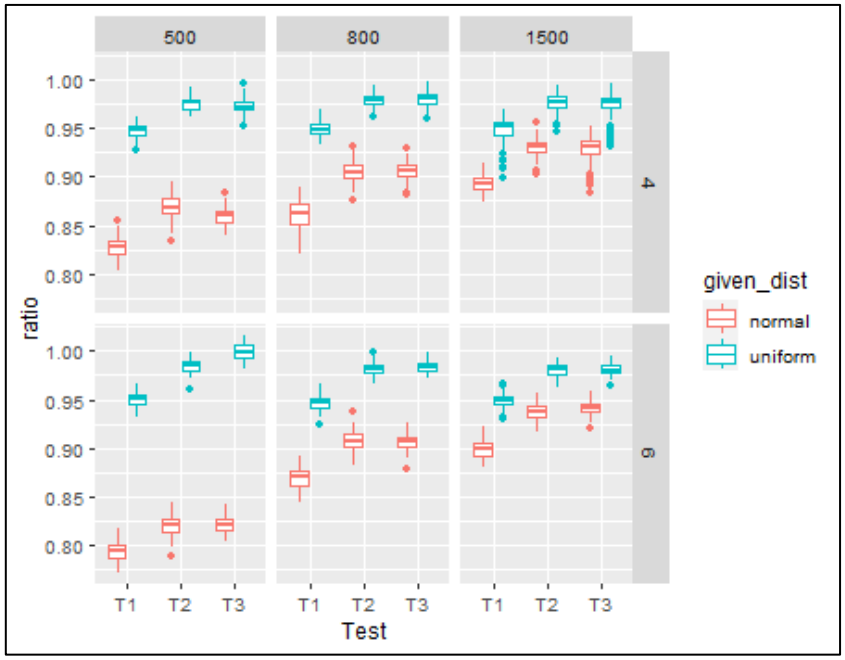


Figure 5
Correlation Index Across Tests and Conditions

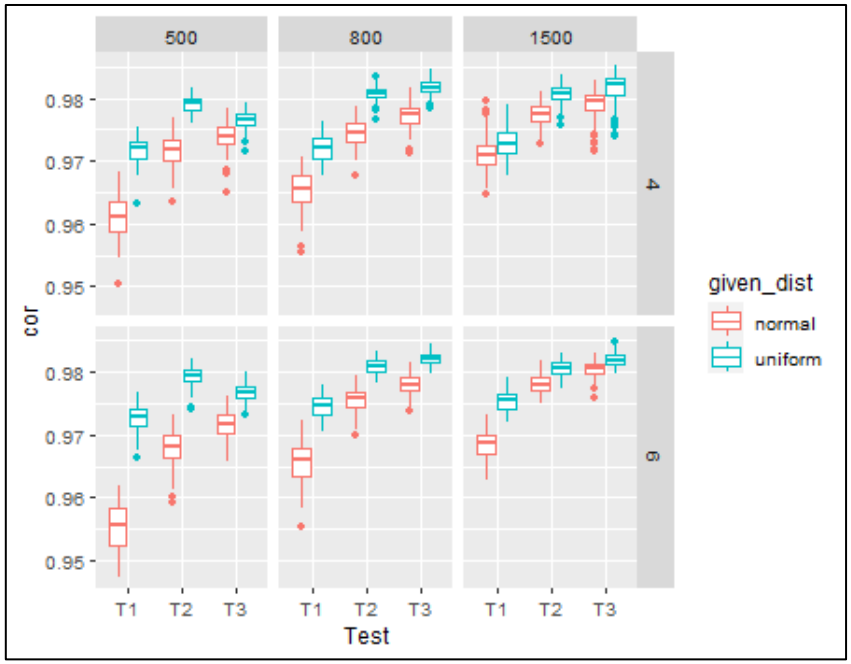


Figure 6
Effects of Item Bank Size on θ Recovery (Mean Across Conditions)

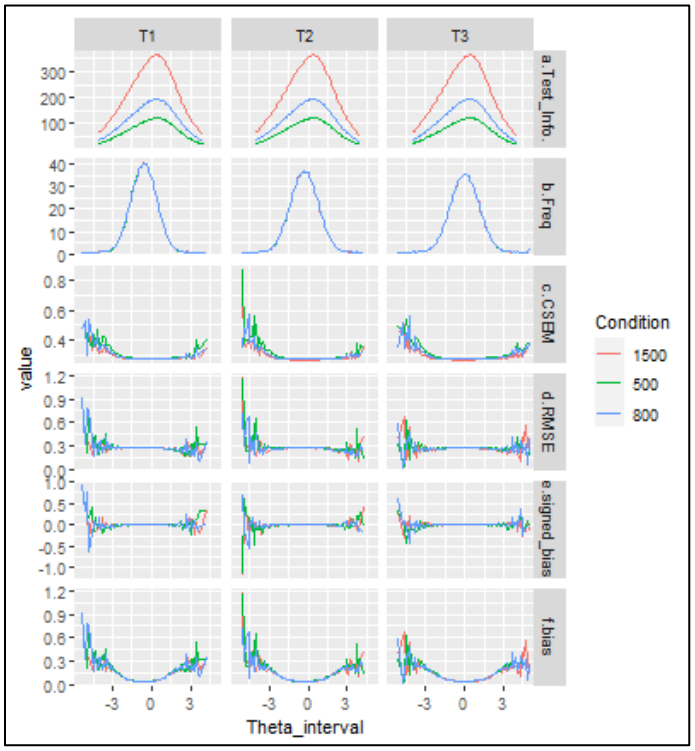


Figure 7
Effects of Item Bank Shape on θ Recovery (Mean Across Conditions)

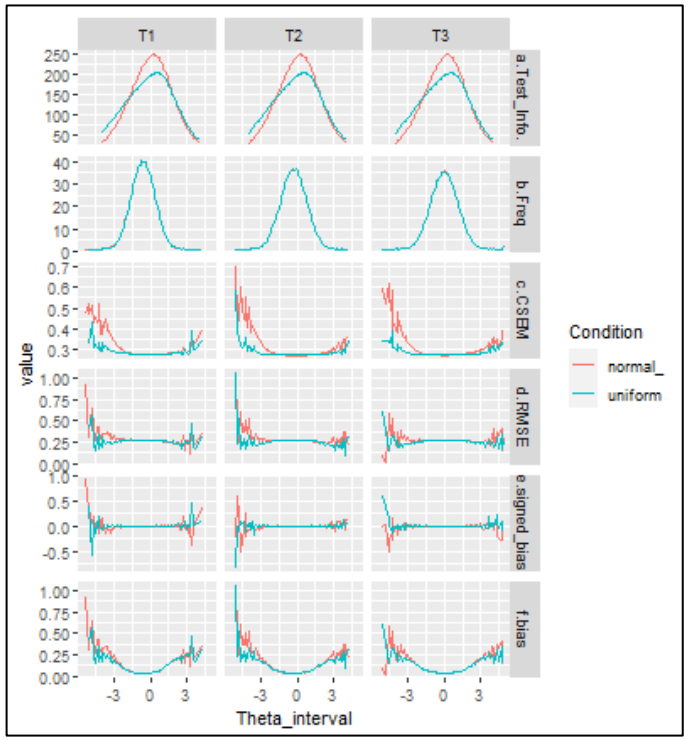


Table 8
Number of Unique Item Exposures by Frequency Category

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.

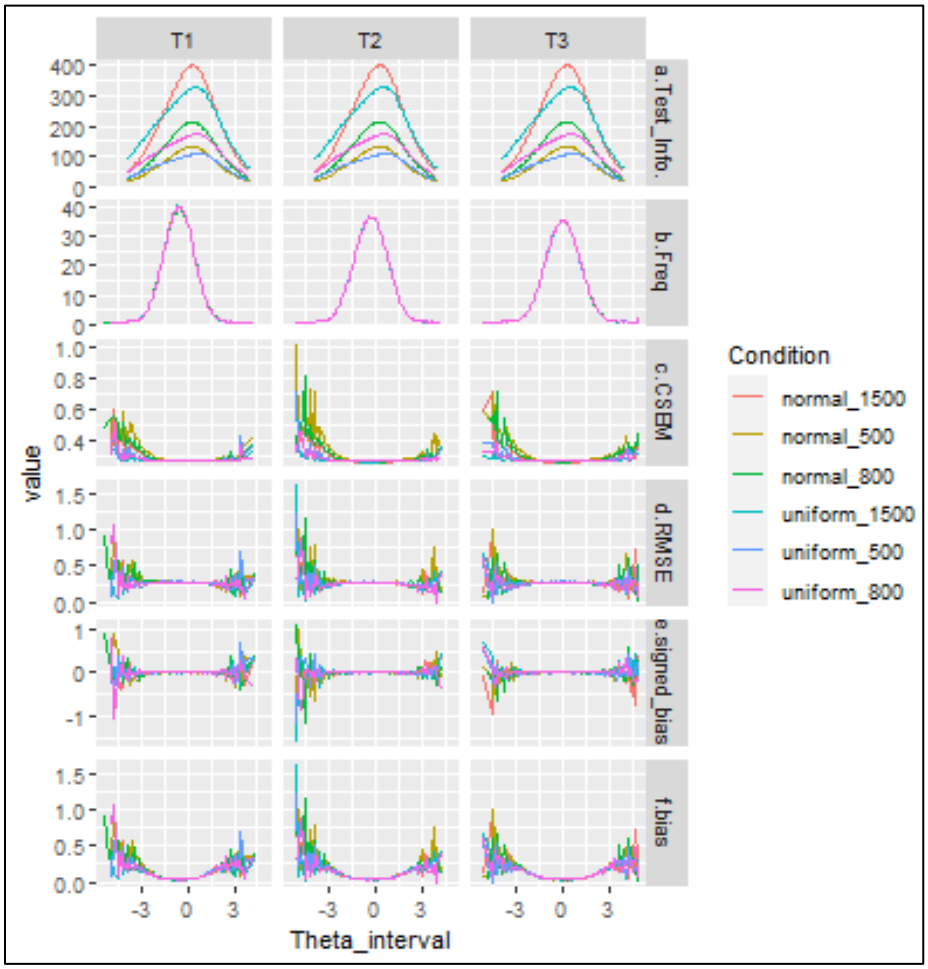
The effects of item bank distribution on on-grade item exposure were mostly inconsequential. The differences in utilization rates between the normal and uniform distributions were negligible for bank sizes of 500, as shown in Table 8. The uniform banks of the 800 and 1,500 items showed slightly higher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

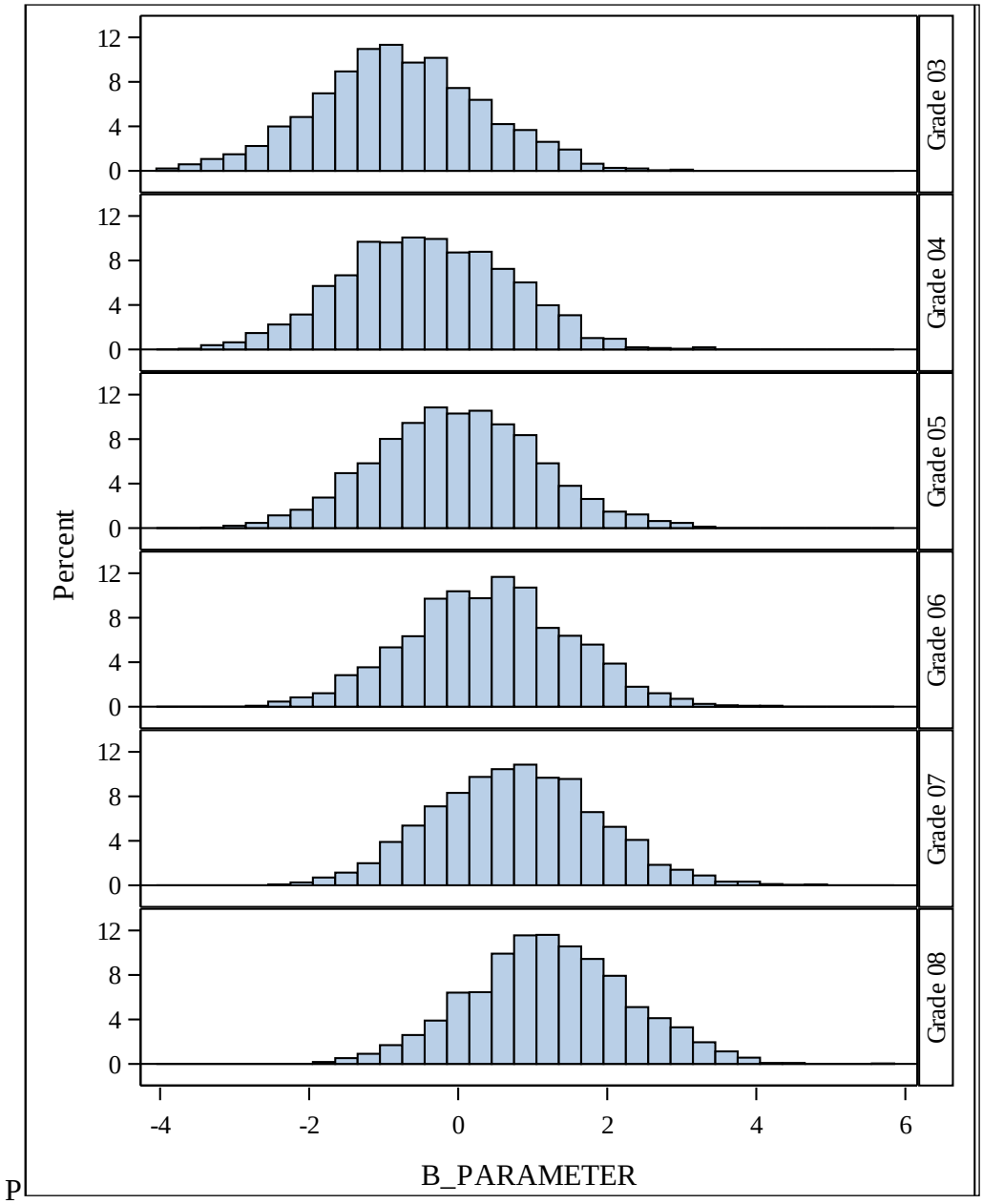
garron@gianopulos.com

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*

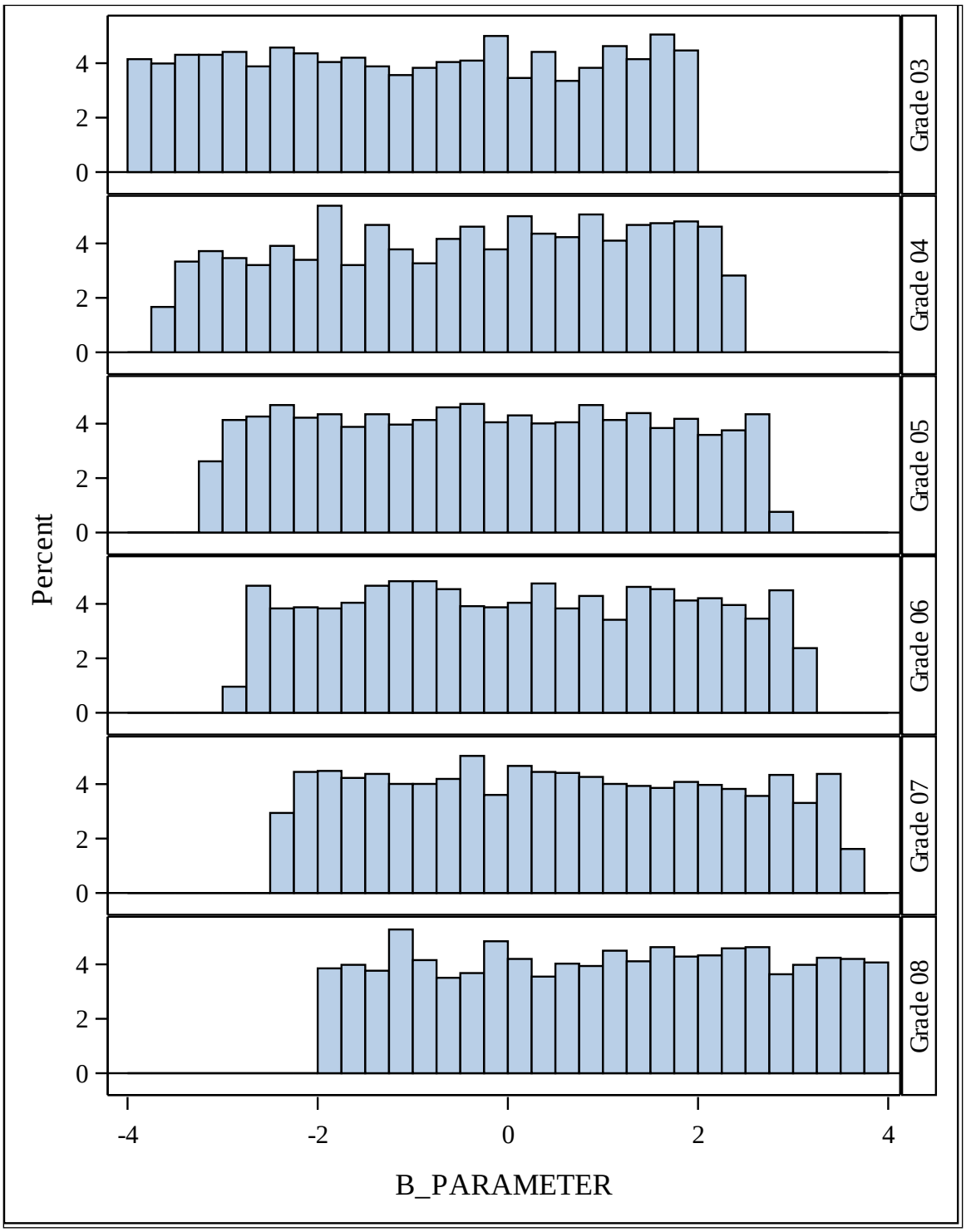
Appendix A: Simulated Item Bank Characteristics

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			

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 10.7333/2502-1201088
ISSN 2165-6592

The Impact of Item Bank Transition Rules on Student Ability Estimates and Achievement Level Classifications

Jonghwan Lee[1], Sangdon Lim[2], M. Christina Schneider[1], Garron Gianopulos[1], Luping Niu[2], Sooyong Lee[2], and Seung W. Choi[2]

[1]NWEA

[2]The University of Texas at Austin

This paper describes a hybrid interim-summative computerized adaptive assessment design administered across three academic terms (fall, winter, and spring). Each test event had two phases that either stayed on-grade or moved off-grade dynamically. Using three different transition rule conditions, this design was compared to an on-grade hybrid interim-summative computerized adaptive assessment design using a series of simulations for a Grade 4 and Grade 6 mathematics assessment. A 500- and 800-item bank in each grade was simulated with a normal distribution of items that ranged across achievement levels. Simulees were randomly drawn from a normal distribution. Simulees whose ability estimates met the transition rule requirements were routed to off-grade item banks and test blueprints dynamically. During Phase 1, the algorithm was programmed to administer only on-grade items. At the end of Phase 1, a transition rule was used to determine if the student should be routed to off-grade items and blueprints for Phase 2. If the student ability estimate did not meet the transition rule requirements, the adaptive test continued to administer on-grade items. If the ability estimate met the transition rule requirements, the student was routed to item banks and blueprints at the adjacent lower or higher grade, as appropriate. The results were evaluated based on measurement precision (i.e., RMSE), measurement accuracy (i.e., bias), item exposure rates, and classification accuracy. Results indicated that while more students were routed off-grade when more lenient routing rules were used, most

of the resulting evaluation indexes remained similar across all conditions, except for item exposure rates that varied across conditions. More lenient transition rules led to fewer overexposed items. More stringent transition rules maintained continuity with the on-grade achievement level designations. This preliminary evidence indicates that it might be feasible to transition students at the tails of the distributions to an off-grade bank and meet federal requirements. Such a transition integrates a fundamental aspect of interim assessments—going off-grade—with the summative assessment ESSA requirement that proficiency determinations be derived from on-grade items and constraints.

Keywords: *classification accuracy, computerized adaptive tests, multi-phase, off-grade, through-year assessments, transition rules*

A new assessment design that connects interim and summative assessment purposes is beginning to gain momentum in K–12 educational assessment. The United States Department of Education (USED) gave flexibility for different assessment system designs in the final regulations for the Every Student Succeeds Act (ESSA) Volume 81 (USED, 2016) by noting the following:

> States have flexibility to develop new assessment designs, which may include a series of multiple statewide interim assessments during the course of the academic year that result in a single summative assessment score (sometimes described as "modular" assessments (p. 3).

Hybrid interim-summative assessments that result in a single summative score are currently referred to as through-year assessments in the field, and states such as Florida, Nebraska, and Texas began piloting or implementing such designs during the 2022–23 school year. In such an approach, typically scores from the third administration are used for federal accountability.

States intend a through-year assessment to provide a more cohesive depiction of how students develop in a state's standards using the summative assessment construct of proficiency throughout the school year as a tool for progress monitoring. These assessments are typically characterized as maintaining the same domain-based blueprint throughout the year across grades and across administrations. Proponents of such systems want teachers to have actionable information that indicates a student's current level of performance and change in performance (growth) across three testing windows to improve classroom teaching and support student learning year-round.

A puzzle that developers of such assessment systems must solve is how to allow off-grade adaptivity, if this is a policy desire, while measuring student performance with on-grade items for accountability purposes. Interim assessment providers have historically allowed students to receive off-grade content because they argued it allows for more precise measurement of student ability, especially at the low end of the scale (Li & Meyer, 2019). Rambo-Hernandez and Warne (2015) argued that a floor and ceiling effect on low- and high-ability students leads to more measurement error when the item bank does not cover a wide range of student ability. They showed that measurement error is minimized when an assessment includes items aligned to students' ability level, indicating that off-grade testing is a solution for providing higher measurement precision for the low- and high-ability students. Way et al. (2010) noted that using some off-grade items is appropriate when a purpose of the assessment is to measure growth in student knowledge and skills across time. Way et al. argued that such an approach is sensible both from a measurement

perspective, and the use of off-grade items aligns to a philosophy of personalized learning. Earlier grade standards are meant to be building blocks to grade-level standards.

Researchers who work in gifted education might have most investigated and advocated for off-grade testing (Achter et al., 1996; Mills & Barnett, 1992; Rambo-Hernandez & Warne, 2015; Stanley & Benbow, 1981; Stanley, 2005; Terman, 1926; VanTassel-Baska et al., 1996). Lohman and Korb (2006) argued that most standardized tests have poor measurement of students who are in the top and bottom decile of ability level and measurement error can lead to inappropriate classroom instruction and decreased student motivation to learn. The research that exists regarding students routing to different paths mainly appears in a multistage testing framework and shows that a great deal of consideration must go into determining the transition rules to achieve maximum efficiency and classification accuracy (Hendrickson, 2007; Svetina et al., 2019; Yan et al., 2016). Recently, Meyer et al. (2023) found that an adaptive algorithm that allows selecting more items at the student's ability estimate from the grade in which the student is enrolled tends to provide a higher final ability estimate than an algorithm that selects items based solely on the student's ability estimate when using the same blueprint but with less focus on selecting items from the grade in which a student is enrolled. Their findings suggest that stakeholders might need to architect careful rules regarding which students receive off-grade items.

Bejar (2016) recommended criterion-referenced routing as a transition rule in the context of summative educational assessments. He suggested that a proficient cutscore could be used as the indicator of how to route a student. Using his suggested model, a student ability estimate could be compared to the proficient level cutscore to determine whether a student should move to a more difficult form or an easier one in the context of a multistage assessment. The Smarter Balanced Assessment Consortium (American Institutes for Research, AIR, 2016) permitted students to be administered off-grade items in a computerized adaptive assessment after the student completed two-thirds of a test event. The transition rule estimated the likelihood a student could be (1) found to be proficient with the below-grade items included in the student's final ability estimate, or (2) denied being recognized as proficient by including above-grade items in the student's final ability estimate. When a criterion-referenced routing approach is applied to an assessment that uses both off-grade and on-grade item banks, such an approach could be used to ensure control regarding which students have access to off-grade items.

Schneider et al. (this volume) described a through-year assessment system prototype that stakeholders conceptualized utilizing criterion-referenced transition rules that serve the following high-level policy goals for a Grade 3–8 assessment system after examining the ESSA regulations (USED, 2016):

1. Summative assessment score interpretations of student ability should underpin the assessment system such that students are able to show comparable ability estimates during different academic terms (fall, winter, and spring). This allows students who meet the cut-score for proficiency to be found in each academic term.
2. Students should have the ability to preserve advanced proficiency so they can move to the next higher adjacent grade when they are individually ready. This means that an advanced student's score for accountability is based on their on-grade responses earlier in the year, whereas the spring administration represents most students' scores for accountability purposes.
3. Students should have multiple opportunities to demonstrate on-grade mastery. Therefore, lower-performing students routed to below-grade items in an academic term should start

each new academic season with on-grade items to quickly determine if those students are now able to access grade-level standards and stay in the item bank of their designated grade. This is done because ESSA regulations and precedent (AIR, 2016) require (1) student proficiency level designations to be based on items administered on-grade, and (2) the proficiency level designation must be made before moving students off-grade to provide supplemental information for teachers. This approach also enhances communication to policymakers regarding which students in a grade are most in need of substantial academic intervention during the year.

The purpose of this paper is to investigate the functionality of the through-year assessment design described by Schneider et al. (this volume).

Research Questions

This study investigated item bank transition rules. The main research questions were:

1. What impact do different transition rule conditions have on the accuracy of student ability estimates for each academic term when used to control if and how students should be administered off-grade items?
2. What impact does allowing off-grade items have on the overall accuracy of student ability estimates for each academic term compared to student ability estimates using only on-grade items?
3. What impact do different off-grade item transition rule conditions have on item exposure and utilization rates compared to when students are only administered on-grade items?

Method

The research questions were examined in a monte-carlo simulation study. Simulated data were used instead of real data for two reasons. First, simulated data separate the effect of model misfit and calibration errors (Bolt, 1999; Davey et al., 1997). Second, stakeholders often desire to see the functioning of new test designs prior to implementation in a pilot.

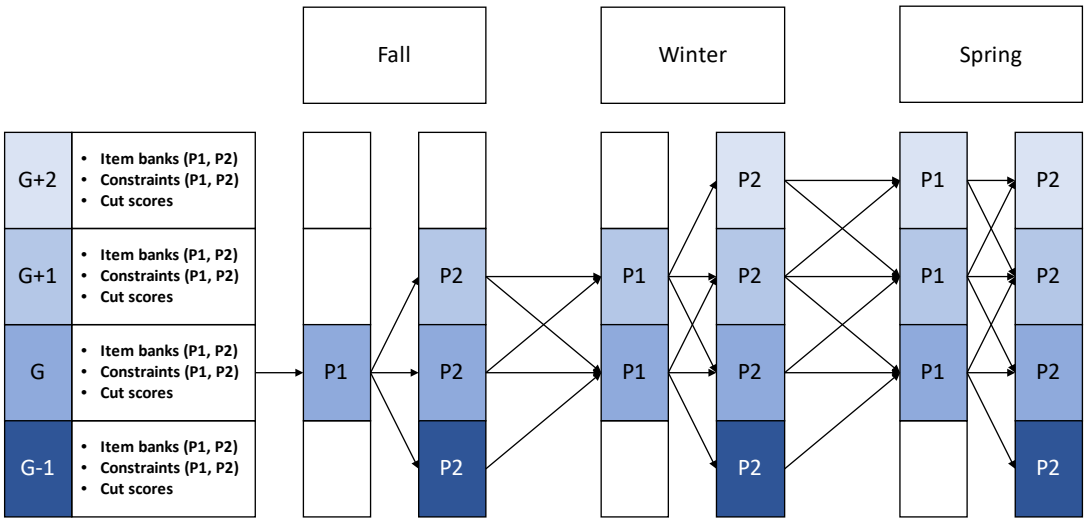
Test Design

A through-year assessment system has the assumption of being administered three times a year in each academic season (i.e., once in the fall, winter, and spring). The through-year adaptive assessment algorithm we investigated comprised two phases for each test event. Each phase included a grade-specific item bank and set of blueprint constraints. This design differed from the typical multistage adaptive assessment that presents each module within a phase as a fixed form, preassembled prior to testing at different levels of difficulty. In the fall, Phase 1 comprised 25 adaptively selected on-grade items for the student's grade (G) of record that determined a student's path in Phase 2. The student ability (θ) estimate and transition rule at the end of Phase 1 determined whether the student should be routed to an off-grade item bank. Phase 2 comprised sequestered item banks and blueprint constraints associated with a particular grade level that adaptively selected 16 items. At the conclusion of Phase 2, responses to all 41 administered items across phases for the fall test event were used to calculate the student's final θ estimate that was then used

to route the student to the appropriate item bank and constraints for Phase 1 of the winter assessment. For this study, the final θ estimate for a student in Phase 2 of each academic season was considered the final ability estimate used to evaluate the performance of the proposed transition rules across time.

In the test design, should a student meet the transition rule criteria, they could be routed to an adjacent grade's item bank and blueprint at the end of each phase. Students who were routed to an adjacent higher grade were constrained to go no farther than two grades above their grade of record ($G + 2$) across all three academic seasons. Students who were routed to a below-grade bank ($G - 1$) during Phase 2 in an academic season were always returned to the on-grade (G) item bank and blueprint constraints for Phase 1 of the next academic season following ESSA regulations (USED, 2016). Thus, the grade range possible across all academic seasons for a particular grade of record was limited from $G - 1$ to $G + 2$, as shown in Figure 1.

Figure 1
Routing of Phases



Source: https://cran.r-project.org/web/packages/maat/vignettes/maat.html

Figure 1 shows three assessment academic seasons: fall, winter, and spring. The assessment within each academic season had two computerized adaptive phases administered as a single test event. The shading in Figure 1 is used to show the grade-level bank and constraints possible for each phase. The arrows show the possible pathways to item banks and blueprint constraints based on the transition rules. For example, the fall has three possible pathways:

1. The on-grade-level Phase 1 paired with an above-grade-level Phase 2.
2. The on-grade-level Phase 1 paired with an on-grade-level Phase 2.
3. The on-grade-level Phase 1 paired with a lower-grade-level Phase 2.

The arrows between the fall and winter administrations show the possible pathways to the bank and constraints that begin Phase 1 of the winter administration, which depends on the student's final θ estimate in the fall and the invoked transition rule.

Simulated Item Banks

Items were simulated to range in difficulty across achievement levels using a normal distribution and included content features found in an existing operational mathematics program used for state accountability purposes. The simulated content included features such as multiple-choice items and technology-enhanced items (e.g., multiple-choice, composite, gap match, graphic gap match, hot text, and text entry), aligned to standards and the corresponding domain, and the use of Webb's (2005) depth of knowledge (DOK). The initial bank size was simulated to be 800 items, but 500 items were also randomly drawn from the original 800 to help gauge the effect on the design functionality in case items must be removed from the bank, which occurs in operational testing programs. Stocking (1988) noted the importance of maintaining the content and statistical characteristics of an item bank as changes are made to ensure that resulting student ability estimates remain comparable. As noted by Schneider et al. (this volume), the simulated item banks were constructed with the assumption that items were aligned to range achievement level descriptors (ALDs; Egan et al., 2012) in sufficient numbers for each achievement level bin. This would allow (1) the blueprints to be met in all achievement levels, and (2) most students in a grade to remain in the grade-level item bank and demonstrate growth by moving into adjacent, higher achievement levels as their ability increased. This approach to creating a sufficiently deep bank was intended to allow most students to remain in the grade-level bank.

Content Constraints

The content constraints used in this study were adapted from an existing Grade 3–8 operational mathematics program used for state accountability purposes. Appendix A presents the specific content constraints for all grades and phases. For each grade, the same proportional blueprint and content constraints were provided for each phase of the test event, resulting in 41 adaptively selected items. For example,

1. Grade G + 1 Winter Phase 1 and Grade G + 1 Winter Phase 2 differed in the number of items required to satisfy a constraint, but the proportional representation to the state blueprint was the same.
2. The Grade G + 1 Winter Phase 1 and Grade G + 1 Winter Phase 2 constraints produced the same overall G + 1 blueprint to the existing operational state mathematics program.
3. Similarly, Grade G Fall Phase 2, Grade G Winter Phase 2, and Grade G Spring Phase 2 used the same constraint configuration.
4. The Grade G Phase 1 and Grade G Phase 2 constraints produced the same overall Grade G blueprint to the existing operational state mathematics program for each academic season.

Simulation Procedures

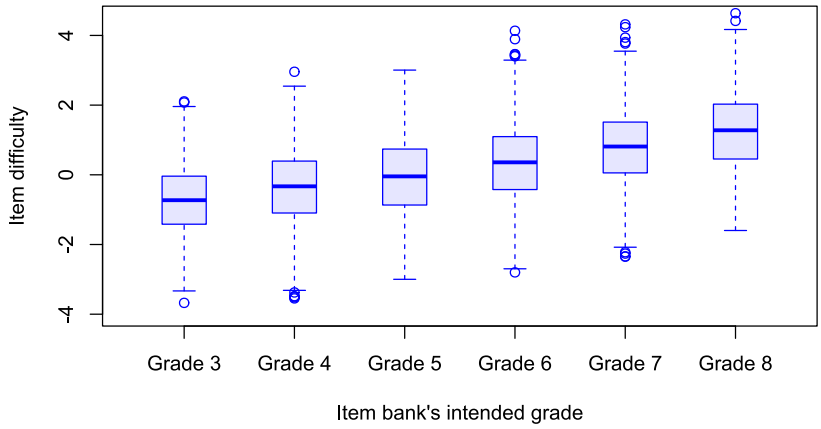
Item parameters. This study used a simulated mathematics item bank for Grades 3–8 with metadata representative of a summative assessment construct (i.e., the inclusion of technology-enhanced items worth more than one point). The item bank size was varied in two levels: 500 and 800. The item banks were generated under a Rasch model (Rasch, 1960) for dichotomous items and a partial credit model (Masters, 1982) for polytomous items.

Table 1 presents the mean and standard deviation (SD) used to generate the item parameters for Grades 3–8. To ensure that the vertical scale was articulated across grades, a lower boundary was set for each grade. 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, as shown in Figure 2. The vertical articulation of item parameters was an assumption of this study's test design, and it contributed to one of the transition rules discussed in the next section.

Table 1
Mean and SD for Normal Distribution

Grade	Mean	SD	Lowest *b* Parameter
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

Figure 2
Vertically Articulated Item Banks Across Grades



Transition rules. Researchers investigated two categories of transition rules to define if and when students were administered off-grade items and blueprints: student-centered and content-centered.

*Student-centered approach***.** The student-centered transition rule was based on using confidence intervals (CIs; Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000) from maximum likelihood estimation (MLE) scoring extended to multiple cutscores (Thompson, 2007). The student-centered approach used CIs obtained from the estimated θ ($\hat{\theta}$) and standard errors from each phase. A CI indicates the range of possible scores where an unknown true score might fall. For example, a 95% CI means that 95 out of 100 times, the true score falls within the range defined by the interval. The comparison between the CIs and pre-specified cutscores of grades determined

routing for the next phase. If the upper bound of a CI was below the lowest on-grade cutscore at the end of Phase 1, the routing would be to the below-grade item bank. If the lower bound of a CI was above the highest on-grade cutscore, the routing would be to the above-grade item bank and blueprint. In all other cases, students stayed on-grade. In the current study, three CI ranges were included: 1.00 conditional standard error of measurement (CSEM) (68% CI), 1.64 CSEM (90% CI), and 1.96 CSEM (95% CI).

Content-centered approach. The content-centered transition rule used characteristics of the item bank to determine the student pathway. The content-centered approach identified the student as needing on- or off-grade items by comparing the Phase 1 $\widehat{\theta}$ to a cutscore representing the ceiling and floor of the within-grade item bank. If the student's $\widehat{\theta}$ was either above the 95th percentile of item difficulty or below the 5th percentile of item difficulty in the Phase 1 item bank, the student was routed to the above- or below-grade item bank in the next phase. In subsequent tables and figures, this approach is denoted as a bank-based transition rule.

Based on these two approaches, the following transition rules were evaluated and compared during the simulations in this study:

1. No transition,
2. CI 68,
3. CI 90,
4. CI 95,
5. Bank-based.

True ability distribution. In each replication, simulees' true θ values were randomly drawn from a normal distribution. Because the three academic season administrations typically occur in fall, winter, and spring, real-world student abilities (on average) tend to increase over time on an interim assessment (NWEA, 2019). The means for fall were set to start at –1.0, and –0.2 for Grades 4 and 6, respectively. The means were set to increase by 0.3 in each subsequent administration. The standard deviations were set to 1. Correlations were set to 0.9 for adjacent seasons and 0.8 for sub-adjacent seasons, similar to values reported for a mathematics interim assessment (NWEA, 2019).

Each replication included 1,000 simulees, and each simulee had three true θ values (one for each academic season). Because the data generation mechanism for each true θ was based on a three-variate distribution with between administration correlations, it was not necessarily the case that simulee θs were generated to be always monotonically increasing over academic seasons.

Exposure and Overlap Control

Bank-based exposure control was not implemented in the simulations (i.e., items were not removed as they became more exposed across students). This is consistent with common practice in K–12 educational assessment computerized adaptive assessments. However, overlap control was used across all six phases. This means that a given student should not receive the same item more than once across academic seasons. This is also consistent with common practice in K–12 educational assessment computerized adaptive assessments, when possible. Overlap control was implemented as a soft constraint, penalizing the item information of previously administered items by M = 100. For example, if the original item information to be used for adaptive test assembly was 5.0 for an item at an interim ability estimate, and if that item has been previously administered,

the information value was modified to be 5.0 – M = −95.0 for the purpose of adaptive test assembly. This means that previously administered items were technically allowed to be selected, if necessary, to meet blueprint requirements. The anticipation for the item bank was that it was sufficiently deep compared to the number of items that needed to be administered to the student population. The goal was to see each item administered in a relatively small percentage of the test administrations (AIR, 2016).

Adaptive Test Assembly

The adaptive form assembly in each phase was performed using an optimal test design approach with shadow tests (van der Linden & Reese, 1998). The optimal test design approach has an advantage of ensuring that all content requirements are strictly met. The item with the maximum information at the current interim θ estimate was selected to be administered to the simulee.

Ability estimation. Interim and final ability estimates were obtained using MLE, with expected a posteriori (EAP) estimation as a fallback for when MLE was not feasible because of extreme responses (e.g., when all item scores were 0). For the purpose of evaluating classification performance, estimated θs were converted into four achievement level categories using a predefined set of cutscores (Table 2) that approximated those found in an operational testing program.

Based on the test design, Phase 2 $\hat{\theta}$s within each academic season were obtained by combining Phase 1 and Phase 2 responses to ensure that all students received scores similar in measurement precision.

Table 2
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

Performance Evaluation

Estimation of accuracy and precision. For each academic season, final θ estimates were used to evaluate the ability estimation performance of evaluated conditions. Root mean square error (RMSE) and bias were calculated as performance measures. A reliability measure was also calculated. Traditional reliability coefficients from classical test theory consider individual items and depend on all examinees to take common items, whereas students receive different items in an adaptive assessment. Therefore, the marginal reliability was calculated (Samejima, 1994) as

$$\text{reliability} = \frac{\sigma_T^2}{\sigma_X^2} \tag{1}$$

where σ_T^2 is the variance of true θs (i.e., $\sigma_T^2 = 1$), and σ_X^2 is the variance of estimated θs.

Classification accuracy. Two measures of classification accuracy were used to evaluate the results: accuracy and Cohen's weighted kappa. Accuracy was computed as the proportion of simulees that had the same achievement level categories between true and estimated θs. Cohen's weighted kappa (1968) was implemented with quadratic weights applied.

Item utilization and exposure rate. The item bank utilization rate was calculated as the number of unique items administered at least once, divided by the number of items in each bank. Item exposure rate was calculated as the number of times an item was administered divided by the number of simulees. This was 1.0 if the item was given to all simulees. A lower rate indicates that the item was not overly exposed to simulees during the test. The exposure rate calculation was based on approaches from AIR (2016) and van der Linden (2003).

Software. The *MAAT* R package (Choi et al., 2022), which implements a multiple administration adaptive testing design, was used to conduct the simulations. *MAAT* is an extension of the R computerized adaptive testing package *TestDesign* (Choi et al., 2021) which performs the adaptive form assembly for each phase in *MAAT*.

Results

Estimation Accuracy and Precision

Table 3 presents the marginal reliability, RMSE, and bias for each condition for fall, winter, and spring. The test-based marginal reliability coefficients ranged from 0.91 to 0.93 across academic seasons across conditions, regardless of bank size. RMSE was 0.27 on average, and the θ bias estimates were near 0 across all conditions. Table 4 presents the reliability of the θ estimates at the end of each phase. The Phase 2 reliabilities represent the estimate based on the 41 administered items. The Phase 1 reliabilities represent the estimate at the time the transition rule decision was made. The Phase 1 marginal reliability coefficients ranged from 0.87 to 0.89 across academic seasons across conditions, regardless of bank size.

Classification Accuracy

Table 5 shows the classification accuracy measures (accuracy and quadratic-weighted kappa) across academic seasons and conditions. The classification measures remained similar across transition rule conditions. This indicates that student routing does not result materially in a loss of accuracy across conditions if using off-grade items and that the transition rules are functioning as intended. Across academic seasons, accuracy decreased slightly (on average 0.84, 0.82, and 0.81 for each season, respectively), whereas weighted kappa increased slightly (0.89–0.90 for each season, respectively). Classifications are most critical to investigate for students who were routed to a below-grade item bank. Appendix B shows the final achievement level designation of students routed to lower-grade items for a simulation. As shown in Appendix B, the more stringent transition rules, CI-95 and bank-based, were preferable to maintain consistency with Phase 1 account-ability classifications.

Table 3
Reliability, RMSE, and Bias by Academic Season

Item Bank Size	Grade	Routing Rule	Fall			Winter			Spring		
			Reliability	RMSE	Bias	Reliability	RMSE	Bias	Reliability	RMSE	Bias
800	4	CI-68	0.928	0.273	-0.001	0.931	0.272	0.000	0.929	0.272	0.000
		CI-90	0.928	0.272	0.000	0.929	0.272	0.000	0.932	0.272	-0.001
		CI-95	0.927	0.271	0.000	0.927	0.272	0.002	0.929	0.272	-0.001
		Bank-based	0.928	0.271	0.002	0.931	0.274	-0.001	0.928	0.274	-0.002
		No routing	0.925	0.273	0.001	0.928	0.274	-0.001	0.922	0.277	0.000
800	6	CI-68	0.927	0.272	0.000	0.931	0.271	-0.001	0.932	0.272	0.001
		CI-90	0.927	0.272	0.000	0.929	0.270	0.000	0.930	0.272	0.000
		CI-95	0.927	0.272	-0.001	0.930	0.272	0.001	0.928	0.272	0.001
		Bank-based	0.926	0.273	-0.001	0.931	0.272	-0.001	0.928	0.273	0.000
		No routing	0.925	0.273	-0.002	0.928	0.273	0.000	0.921	0.275	-0.002
500	4	CI-68	0.925	0.273	0.002	0.930	0.274	0.000	0.929	0.275	-0.001
		CI-90	0.925	0.273	0.002	0.929	0.274	0.000	0.928	0.276	-0.001
		CI-95	0.926	0.273	0.002	0.929	0.275	-0.001	0.926	0.277	0.000
		Bank-based	0.927	0.273	0.002	0.929	0.274	0.000	0.928	0.277	0.001
		No routing	0.922	0.274	0.001	0.919	0.277	0.000	0.914	0.283	-0.001
500	6	CI-68	0.921	0.275	0.000	0.926	0.274	0.000	0.927	0.276	-0.002
		CI-90	0.921	0.275	-0.001	0.925	0.274	-0.001	0.929	0.277	0.000
		CI-95	0.922	0.275	-0.002	0.925	0.276	-0.002	0.924	0.276	0.001
		Bank-based	0.922	0.275	-0.002	0.928	0.277	-0.001	0.923	0.278	-0.001
		No routing	0.920	0.276	-0.002	0.916	0.280	-0.001	0.910	0.285	-0.002

Note. CI-68 = 68% CI; CI-90 = 90% CI; CI-95 = 95% CI; Bank-based = 5th and 95th item bank difficulty; No routing = on-grade item bank only.

Table 4
Reliability Estimate at the End of Each Phase

Item Bank Size	Grade	Routing Rule	Fall		Winter		Spring	
			Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
800	4	CI-68	0.879	0.928	0.888	0.931	0.883	0.929
		CI-90	0.879	0.928	0.883	0.929	0.887	0.932
		CI-95	0.879	0.927	0.884	0.927	0.882	0.929
		Bank-Based	0.879	0.928	0.887	0.931	0.882	0.928
		No routing	0.879	0.925	0.884	0.928	0.878	0.922
800	6	CI-68	0.878	0.927	0.885	0.931	0.886	0.932
		CI-90	0.878	0.927	0.885	0.929	0.885	0.930
		CI-95	0.878	0.927	0.888	0.930	0.883	0.928
		Bank-based	0.878	0.926	0.885	0.931	0.882	0.928
		No routing	0.878	0.925	0.883	0.928	0.879	0.921
500	4	CI-68	0.876	0.925	0.883	0.930	0.881	0.929
		CI-90	0.876	0.925	0.883	0.929	0.879	0.928
		CI-95	0.876	0.926	0.881	0.929	0.876	0.926
		Bank-based	0.876	0.927	0.883	0.929	0.878	0.928
		No routing	0.876	0.922	0.876	0.919	0.869	0.914
500	6	CI-68	0.870	0.921	0.881	0.926	0.879	0.927
		CI-90	0.870	0.921	0.879	0.925	0.880	0.929
		CI-95	0.870	0.922	0.878	0.925	0.875	0.924
		Bank-based	0.870	0.922	0.878	0.928	0.872	0.923
		No routing	0.870	0.920	0.871	0.916	0.866	0.910

Note. Phase 1 consisted of 25 items, and Phase 2 consisted of 16 items. Phase 2 reliability represents the reliability of ability estimates from both Phases combined (41 items).

Table 5
Classification Accuracy

Item Bank Size	Grade	Routing Rule	Fall		Winter		Spring	
			Accuracy	κ_{weighted}	Accuracy	κ_{weighted}	Accuracy	κ_{weighted}
800	4	CI-68	0.841	0.888	0.823	0.893	0.813	0.897
		CI-90	0.842	0.888	0.821	0.893	0.812	0.896
		CI-95	0.841	0.888	0.823	0.894	0.813	0.897
		Bank-based	0.840	0.888	0.822	0.893	0.812	0.896
		No routing	0.840	0.887	0.825	0.895	0.812	0.896
800	6	CI-68	0.840	0.887	0.826	0.895	0.813	0.897
		CI-90	0.841	0.888	0.826	0.895	0.812	0.896
		CI-95	0.841	0.888	0.824	0.894	0.813	0.896
		Bank-based	0.840	0.887	0.823	0.893	0.812	0.896
		No routing	0.839	0.887	0.823	0.893	0.813	0.897
500	4	CI-68	0.841	0.888	0.823	0.893	0.809	0.894
		CI-90	0.842	0.889	0.823	0.893	0.810	0.895
		CI-95	0.841	0.888	0.821	0.893	0.809	0.894
		Bank-based	0.841	0.888	0.821	0.892	0.809	0.894
		No routing	0.843	0.889	0.822	0.893	0.811	0.896
500	6	CI-68	0.840	0.888	0.822	0.893	0.811	0.896
		CI-90	0.840	0.887	0.824	0.895	0.809	0.894
		CI-95	0.839	0.887	0.823	0.893	0.811	0.896
		Bank-based	0.840	0.888	0.821	0.892	0.809	0.894
		No routing	0.840	0.887	0.821	0.893	0.811	0.895

Note. Grade marked with asterisk (*) is the grade of record. Displayed data is from the 800-item bank condition.

Number of Students Moving On- and Off-Grade

Figure 3 presents student routing diagrams to illustrate the number of students being routed to each pathway when off-grade routing is allowed for grade 6 with transition rule CI-68 as example. Each blue box is a test phase, and every two consecutive boxes represent a test administration. For the Fall test, 645 students remained on-grade, 335 students moved below grade, and 20 students moved to the upper grade from Phase 1 to Phase 2. When the fall test completed, 644 students remained on-grade, 1 student moved from on-grade to the upper grade, and the 335 students who had been moved to the below grade blueprint, constraints, and bank were sent back to the on-grade blueprint, constraints, and bank to begin Phase 1 of the next administration. This process repeated until the spring test concluded. Appendix C shows the routing diagram for Grade 4 and Grade 6 simulees for all transition rules. Comparing the transition rules, the CI-based rules with more conservative criteria had fewer students administered off-grade items. The bank-based transition rule showed the smallest number of students administered off-grade items, effectively making it a stricter rule than CI-based rules.

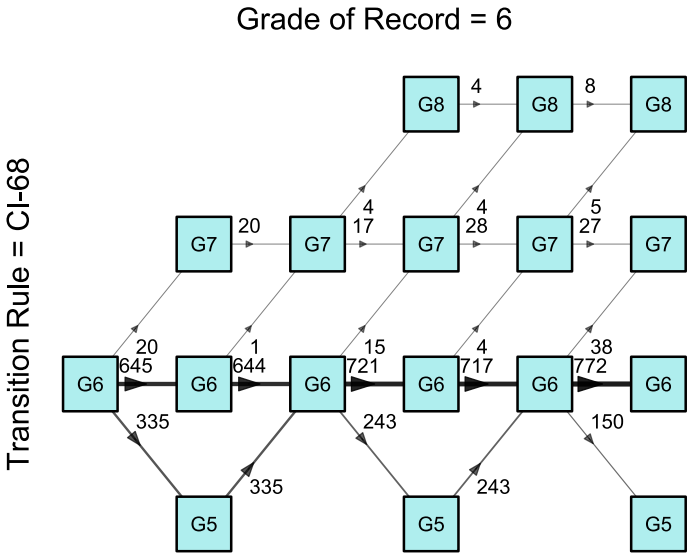
A shared pattern among the routing rules was that as the assessment progressed into later academic seasons, the number of students moving to the below-grade item bank gradually decreased. While this is mainly due to the data generation mechanism in this simulation, where student θs were generated from higher means in later academic seasons, this data generation artifact is based on what is observed in an interim assessment. It is intended in a live student population to capture students' growing abilities in the on-grade standards.

Item Bank Utilization and Exposure Rate

Whether any simulees received an item more than once across phases was also examined. This did not occur. Table 6 presents the item bank utilization rate under different conditions. The value of 1.0 indicates that all items from the bank were utilized. For example, for the Grade 4 bank with the transition rule based on CI-68, the value 0.88 indicates that 88% of the items in the Grade 4 bank were used in the fall simulation. As simulees who met the transition rules were routed to the adjacent off-grade banks in Phase 2, 24% of the Grade 3 items were utilized, whereas only 11% of the Grade 5 items were utilized. The item bank utilization rates differed across transition rules mainly because the number of students who were routed off-grade was dependent on which rule was used. The on-grade item bank utilization rate was not substantively different when transition rules were and were not invoked. The utilization rate for the item bank that stayed on-grade with no transition was 91% versus 88% for the most lenient transition rule. When narrower CI-based transition rules were used (i.e., more lenient routing), higher utilization rates were observed for off-grade banks, as would be expected from more lenient routing rules. The bank-based approach resulted in lower utilization rates for off-grade banks. Combined with results on estimation and classification performance, these results suggest that different routing rules can yield similar levels of reliability while varying item bank utilization.

Figure 4 shows the proportion of items that had specific ranges of exposure rates across conditions. Exposure rates shown in Figure 4 were obtained from items administered throughout all academic seasons combined. In general, conservative transition rules tended to have more

Figure 3
Number of Students Moving On- and Off-Grade



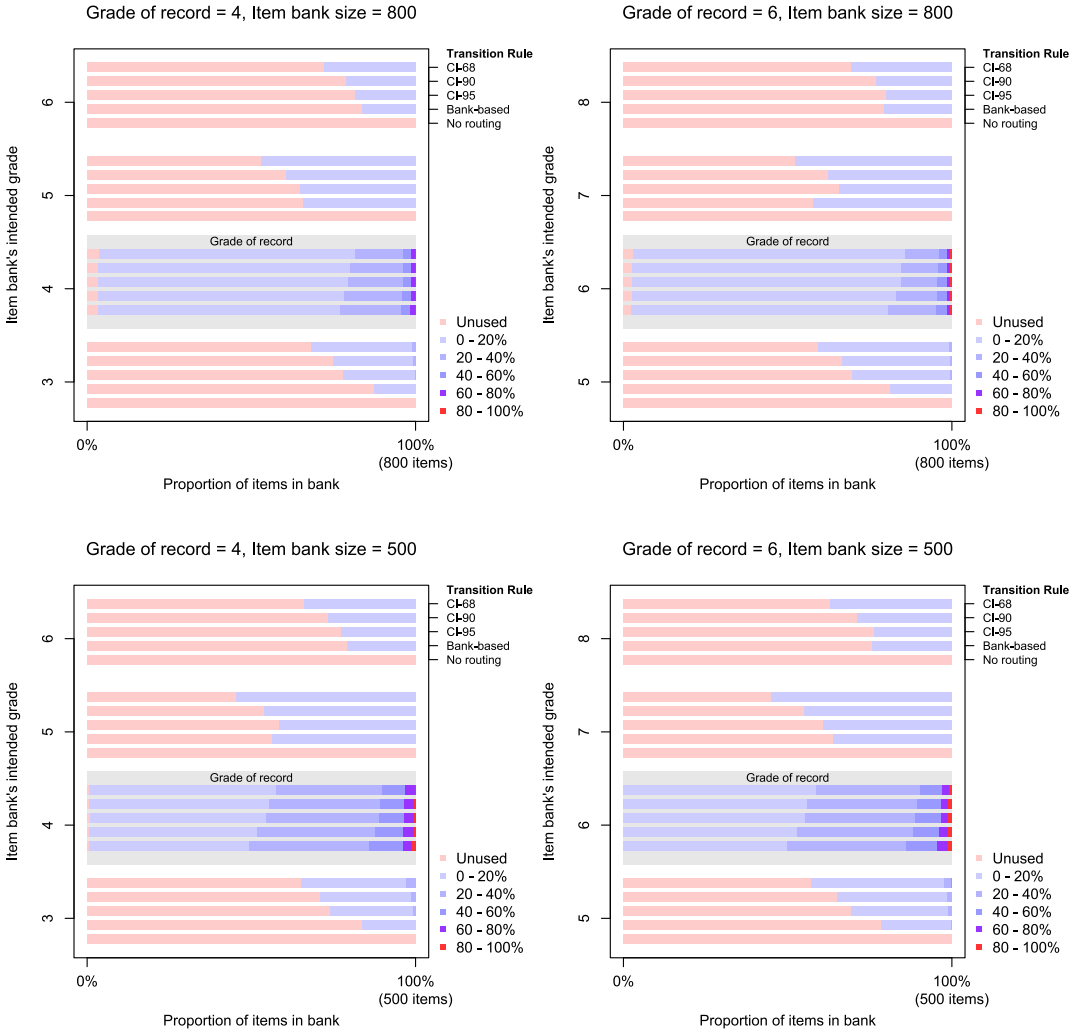
Note. Routing paths that had at least one simulee are displayed.

Table 6
Item Utilization Rates

Routing Rule	Item Bank Grade	Fall	Winter	Spring	Item Bank Grade	Fall	Winter	Spring
CI-68	3	0.242	0.270	0.283	5	0.336	0.358	0.365
	4*	0.884	0.937	0.960	6*	0.894	0.941	0.968
	5	0.109	0.335	0.428	7	0.112	0.326	0.426
	6	0.000	0.071	0.277	8	0.000	0.072	0.300
CI-90	3	0.193	0.210	0.222	5	0.278	0.290	0.296
	4*	0.887	0.941	0.963	6*	0.901	0.946	0.972
	5	0.077	0.263	0.356	7	0.081	0.243	0.335
	6	0.000	0.043	0.208	8	0.000	0.044	0.225
CI-95	3	0.167	0.183	0.193	5	0.256	0.265	0.269
	4*	0.889	0.942	0.964	6*	0.904	0.947	0.973
	5	0.061	0.223	0.315	7	0.067	0.218	0.308
	6	0.000	0.032	0.181	8	0.000	0.037	0.194
Bank-based	3	0.092	0.099	0.110	5	0.154	0.153	0.150
	4*	0.897	0.944	0.965	6*	0.903	0.947	0.972
	5	0.069	0.219	0.306	7	0.099	0.273	0.382
	6	0.000	0.033	0.159	8	0.000	0.049	0.199
No transition	4*	0.905	0.946	0.965	6*	0.911	0.956	0.974

Note. Grades marked asterisk (*) are the grade of record. Displayed data is from the 800-item bank condition.

Figure 4
Exposure Rate Distributions from All Seasons Combined



overexposed items in on-grade banks, as would be expected from conservative routing rules. The bank-based routing rule, being the strictest rule (excluding the no routing condition), had the most overexposed items in on-grade banks. These results suggest that when maintaining adequate item exposure rates is a concern, implementing an off-grade routing rule can provide an increase in items with adequate exposure rates (0%–20%) in the on-grade bank and a decrease in overexposed items (> 20%) in the on-grade bank while maintaining similar levels of θ estimation performance, if this is the primary focus of stakeholders.

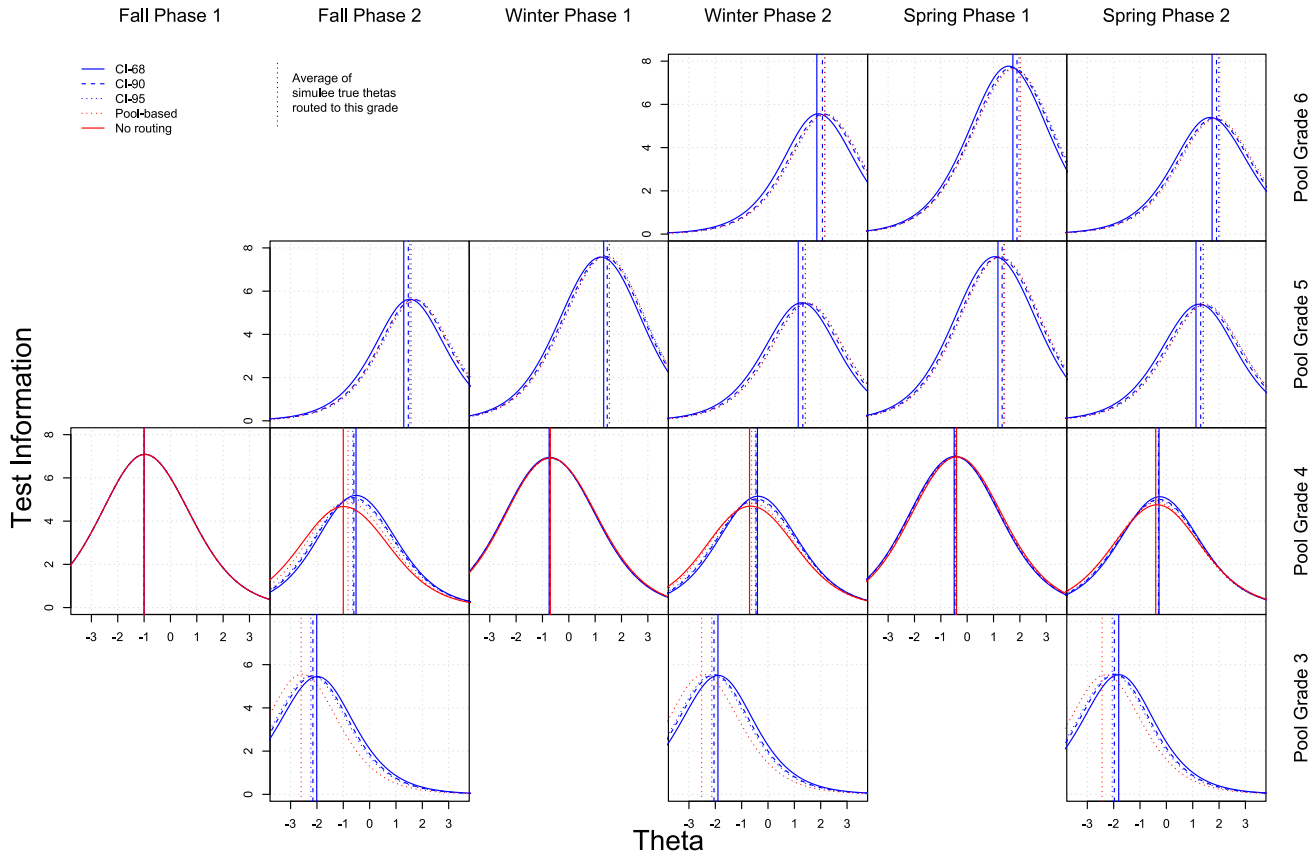
Test Information by Phase

Figure 5 shows the average test information for each phase across academic seasons for simulees who did and did not meet the transition rule criteria across conditions for the 800-item bank. Patterns were similar across the examined conditions, so only the data from the Grade 4 800-item bank condition is displayed. Throughout all six phases and item bank grades, adaptive tests

had close-to-optimal test information at simulees' true θs. This is consistent with performance expected from shadow-test-based optimal test assembly.

There was some variation in test information by routing rule. First, in the on-grade item banks (the row labeled "Bank Grade 4"), using lenient routing rules tended to yield higher test information and using conservative routing rules tended to yield lower test information. One possible cause of this pattern is that when a student stays on-grade in Phase 1 and Phase 2 in a single academic season, the form assembled in Phase 2 is subject to overlap control (i.e., items already administered in Phase 1 should not be included in Phase 2), which would lead to a less informative form in Phase 2. In contrast, when a student is routed off-grade in Phase 2, the form in Phase 2 is allowed to be assembled from the off-grade item bank, which has a more informative set of items for the simulee. Second, more conservative transition rules (i.e., CI-95, bank-based) tended to route students with more extreme θs to off-grade banks, which is consistent with how such routing rules would be expected to behave. This led to adaptive forms in off-grade banks having test information curves that were shifted toward more extreme θs.

Figure 5
Average Test Information Across Seasons by Routing Rule



Note. Test information functions shown in each cell is an average obtained from all adaptive tests administered in that season/phase/grade combination. Displayed data is from the [Grade of Record = 4, Item Bank Size = 800] condition.

Discussion

We investigated (1) the impact different transition rule conditions had on the accuracy of student θ estimates for each academic term when used to control if and how students should be administered off-grade items, and (2) how these results compared to estimates that were outcomes of simulees who were only administered on-grade items. Student-centered transition rules and the content-centered (bank-based) transition rule were compared to a no transition rule condition using 500-item and 800-item banks. Student θ estimates and classification accuracy estimates for achievement levels remained similar across all conditions. This preliminary evidence indicates that it might be feasible to transition students at the tails of the distributions to an off-grade bank, which supports more precise ability estimates. The design approach integrates a fundamental aspect of interim assessments—going off-grade—while meeting the ESSA requirement for proficiency determinations being derived from on-grade items and constraints.

The reliability of θ estimates and bias were sufficiently robust during Phase 1 as to meet criteria for making proficiency determinations. Adequate reliability, when the criterion-referenced transition decision is made (Bejar, 2016) and an assurance that students who are routed to off grade items are neither prohibited from being designated from proficient based on above-grade items nor included as proficient based on below-grade items (AIR, 2016) were among the considerations that needed to be investigated with the design prototype. The results from this simulation study indicated that the design met these criteria. Simulees routed to below-grade items using the more stringent transition rules maintained the same achievement level status when final θ estimates were aggregated across Phase 1 and Phase 2 (see Appendix B). Transitioning select students to off-grade items and communicating that information to teachers in carefully designed score reports should support teachers in better understanding the set of standards in which the student is currently functioning. Lohman and Korb (2006) argued that such information would better serve instructional decision making and student motivation to learn. The impact different conditions had on item exposure and utilization rates were also investigated. Across transition rules (including all students staying in the on-grade item bank), on-grade item utilization rates were similar. As expected, more items were exposed when the bank comprised 500 items than when the bank comprised 800 items. The off-grade item utilization and exposure rates were most similar for CI-90 and CI-95.

The content-centered approach resulted in the lowest off-grade item utilization rates because this transition rule routed the fewest students to off-grade items. The content-centered rule is optimal for states who would want to route students to off-grade items only when student abilities are above or below the on-grade items in the bank. Using such a rule could serve to support the reliability of the fall administration when the on-grade item bank might be too difficult for students early in the year. The simulation results showed that different transition rules lead to variations in the number of students routed to off-grade banks. However, if student ability trajectories increase across academic seasons in the same way as simulated in this study (and similar to what is observed in an interim assessment), always starting Phase 1 with on-grade items for students who were previously routed off-grade should also support reducing the number of students who move off-grade, given the findings of Meyer et al. (2023).

Limitations

As with any simulation study, these findings should be interpreted in context of the item banks used in the simulation. The item banks in the current study were generated to be vertically articulated and to function with state assessment content constraints. These data allowed a wide range of items to be available for the adaptive test assembly across simulees routed to on- and off-grade item banks. It is common for item banks for adaptive assessments to have areas where more items are needed for optimal adaptivity. The way items were simulated likely allowed the current adaptive test assembly simulation to maintain a high degree of reliability across the examined conditions. It would be useful to investigate the design using the item banks and constraints available from an on-grade through-year testing program to examine if the findings differ significantly from the results of this study when transition rules are applied. It would also be wise to administer the design to students in a small-scale pilot to verify the functionality. Which and if transition rules should be implemented is ultimately a state policy decision, but this study does lead to the conclusion that the quality of the item bank is likely a driving factor for ensuring that students are not moved to off-grade items due to areas of sparseness in the item bank.

This simulation modeled student change in ability over time (i.e., student growth), which was implemented as academic season changes in true θs. The design prototype investigated was able to accurately estimate each academic season θs, regardless of whether transition rules were used. Through-year assessments need to investigate if item banks are sufficient to accurately estimate student θs for each academic season. The shadow forms adaptively assembled for each test event had close-to-optimal test information for each academic season true θs (see Figure 5). These results show what might be feasible, but the results are also predicated on the notion that the field can engineer test score interpretations to help teachers understand how students are growing in the complexity and difficulty of the state standards (Schneider et al., this volume). Also of importance was that the data generation mechanism for the academic season true θs did not constrain θs to be monotonically increasing from fall, winter, and spring. Not all students increase in ability across the year. Still, different methods of generating parameters for academic season true θs might lead to different results.

References

Achter, J. A., Lubinski, D., & Benbow, C. P. (1996). Multipotentiality among the intellectually gifted: “It was never there and already it's vanishing.” *Journal of Counseling Psychology, 43*(1), 65–76. *CrossRef*

American Institutes for Research (AIR). (2016, October). *Smarter Balanced summative assessments simulation results*. *WebLink*

Bejar, I. (2016). Past and future of multistage testing in educational reform. In D. Yan, A. A. Von Davier, & C. Lewis, (Eds.), *Computerized multistage testing: Theory and applications*. CRC Press.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12*(4), 383–407. *CrossRef*

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. *CrossRef*

Choi, S. W., Lim, S., Niu, L., & Lee, S. (2022). *MAAT: Multiple administrations adaptive testing*. R package version 1.0.2.9000. *WebLink*

Choi, S. W., Lim, S., & van der Linden, W. J. (2021). TestDesign: An optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*. *CrossRef*

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data*. ACT Research Report Series 97-4. *WebLink*

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

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

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice , 26*(2), 44–52. *CrossRef*

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–254). Academic Press.

Li, S., & Meyer, J. P. (2019). *Simulation study for evaluating MAP® Growth™ grade-level item banks*. NWEA. *Weblink*

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29*(4), 451–484. *CrossRef*

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Meyer, J. P., Hu, A., & Li, S. (2023). *Content proximity spring 2022 pilot study*. NWEA Research Brief. *WebLink*

Mills, C. J., & Barnett, L. B. (1992). The use of the Secondary School Admission Test (SSAT) to identify academically talented elementary school students. *Gifted Child Quarterly, 36*(3), 155–159. *CrossRef*

NWEA. (2019). *MAP® Growth™ technical report*. *WebLink*

Rambo-Hernandez, K. E., & Warne, R. T. (2015). Measuring the outliers: An introduction to out-of-level testing with high-achieving students. *Teaching Exceptional Children, 47*(4), 199–207. *CrossRef*

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

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229–244. *CrossRef*

Schneider, M. C., Choi., S. W., & Lewis, D. (this volume, 2025). Design considerations and reporting solutions for a multiple administrations adaptive testing system. *Journal of Computerized Adaptive Testing, 12(1),* 88-122. *CrossRef*

Stanley, J. C. (2005). A quiet revolution: Finding boys and girls who reason exceptionally well and/or verbally and helping them get the supplemental educational opportunities they need. *High Ability Studies 16*(1), 5–14. *CrossRef*

Stanley, J. C., & Benbow, C. P. (1981). *Using the SAT to find intellectually talented seventh graders*. College Board Review.

Stocking, M. L. (1988, May). *Some considerations in maintaining adaptive test item pools*. Educational Testing Service. *WebLink*

Svetina, D., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement 56*(1), 192–213. *CrossRef*

Terman, L. M. (1926). *Genetic studies of genius...: Vol. I. Mental and Physical Traits of a Thousand Gifted Children*. Stanford University Press.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, and Evaluation*, *12*(1), 1. *CrossRef*

U.S. Department of Education (USED). (2016). *Final regulations* (Docket ID ED-2016-OESE-0032). *WebLink*

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

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *28*(3), 249–265. *CrossRef*

VanTassel-Baska, J., Benbow, C. P., & Lubinski, D. (1996). *Contributions of the talent-search concept to gifted education* (pp. 236–245). Johns Hopkins University Press.

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the common core assessments*. Unpublished manuscript.

Webb, N. L. (2005). *Web Alignment Tool (WAT): Training manual (Draft Version 1.1).* Wisconsin Center for Education Research, Council of Chief State School Officers.

Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2014) *Computerized multistage testing: Theory and applications*. CRC Press.

Acknowledgements and Assistance

The MAAT simulator and design prototype was funded by NWEA.

MAAT is open-source and maintained by Seung Choi at the University of Texas.
We also would like to acknowledge Kelly Rivard for her editorial support.

Author Address

Jonghwan Lee Email: jay.lee1@pearson.com

Citation

Lee. J., Lim, S., Schneider, M.C., Gianopulos, G., Niu, L., Lee., S., & Choi, S.W. (2025).
The impact of item bank transition rules on student ability estimates
and achievement level classifications,
Journal of Computerized Adaptive Testing, 12(1), 88-122.

Appendix A: Constraints

Table A-1
Grade 3 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 3.1.1.a" & DOK %in% c(1, 2)	2–3	1–2
4	Number	STANDARD == "MA 3.1.1.b" & DOK %in% c(1, 1)	0–1	0–1
5	Number	STANDARD == "MA 3.1.1.c" & DOK %in% c(1, 1)	0–1	0–1
6	Number	STANDARD == "MA 3.1.1.d" & DOK %in% c(1, 2)	1–2	1–2
7	Number	STANDARD == "MA 3.1.1.e" & DOK %in% c(1, 2)	0–1	0–1
8	Number	STANDARD == "MA 3.1.1.f" & DOK %in% c(1, 2)	1	1
9	Number	STANDARD == "MA 3.1.1.g" & DOK %in% c(1, 2)	1	1
10	Number	STANDARD == "MA 3.1.1.h" & DOK %in% c(1, 2)	0–1	0–1
11	Number	STANDARD == "MA 3.1.1.i" & DOK %in% c(1, 2)	0–1	0–1
12	Number	DOMAIN == "NR"	5–6	5
13	Number	STANDARD == "MA 3.1.2.a" & DOK %in% c(1, 1)	0–1	0–1
14	Number	STANDARD == "MA 3.1.2.c" & DOK %in% c(1, 2)	0–1	0–1
15	Number	STANDARD == "MA 3.1.2.e" & DOK %in% c(1, 1)	0–1	0–1
16	Number	STANDARD == "MA 3.1.2.f" & DOK %in% c(1, 2)	0–1	0–1
17	Number	DOMAIN == "NO"	3	1–2
18	Number	UDOMAIN == "NUM"	9	7
19	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–9	1–7
20	Number	STANDARD == "MA 3.2.1.a" & DOK %in% c(1, 2)	0–1	0–1
21	Number	STANDARD == "MA 3.2.1.b" & DOK %in% c(1, 2)	0–1	0–1
22	Number	DOMAIN == "AR"	1	0–1
23	Number	STANDARD == "MA 3.2.2.b" & DOK %in% c(1, 2)	1–2	1
24	Number	DOMAIN == "AP"	1–2	1
25	Number	STANDARD == "MA 3.2.3.a" & DOK %in% c(2, 2)	0–1	0–1
26	Number	STANDARD == "MA 3.2.3.b" & DOK %in% c(2, 2)	0–1	0–1
27	Number	DOMAIN == "AA"	1–2	0–1
28	Number	UDOMAIN == "ALG"	4	2
29	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–4	1–2
30	Number	STANDARD == "MA 3.3.1.a" & DOK %in% c(1, 1)	2	1–2
31	Number	STANDARD == "MA 3.3.1.b" & DOK %in% c(1, 2)	0–1	0–1
32	Number	STANDARD == "MA 3.3.1.c" & DOK %in% c(1, 2)	1–2	1
33	Number	DOMAIN == "GC"	3	2
34	Number	STANDARD == "MA 3.3.3.a" & DOK %in% c(1, 2)	0–1	0–1
35	Number	STANDARD == "MA 3.3.3.b" & DOK %in% c(1, 1)	0–1	0–1
36	Number	STANDARD == "MA 3.3.3.c" & DOK %in% c(2, 2)	1	0–1

37	Number	STANDARD == "MA 3.3.3.e" & DOK %in% c(1, 1)	1	1
38	Number	STANDARD == "MA 3.3.3.g" & DOK %in% c(1, 2)	0–1	0–1
39	Number	STANDARD == "MA 3.3.3.h" & DOK %in% c(1, 3)	0–1	0–1
40	Number	DOMAIN == "GM"	4	2
41	Number	UDOMAIN == "GEO"	7	4
42	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–7	1–4
43	Number	STANDARD == "MA 3.4.1.a" & DOK %in% c(2, 3)	2	1–2
44	Number	STANDARD == "MA 3.4.1.b" & DOK %in% c(1, 2)	1	0–1
45	Number	DOMAIN == "DR"	3	2
46	Number	STANDARD == "MA 3.4.2.a" & DOK %in% c(2, 2)	2	1
47	Number	DOMAIN == "DA"	2	1
48	Number	UDOMAIN == "DTA"	5	3
49	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–5	1–3
50	SUM	POINTS	29–32	20–21

Table A-2
Grade 4 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 4.1.1.a" & DOK %in% c(1, 2)	0–1	0–1
4	Number	STANDARD == "MA 4.1.1.c" & DOK %in% c(1, 1)	0–1	0–1
5	Number	STANDARD == "MA 4.1.1.d" & DOK %in% c(1, 1)	0–1	0–1
6	Number	STANDARD == "MA 4.1.1.e" & DOK %in% c(1, 1)	0–1	0–1
7	Number	STANDARD == "MA 4.1.1.f" & DOK %in% c(1, 1)	0–2	0–1
8	Number	STANDARD == "MA 4.1.1.g" & DOK %in% c(1, 1)	0–1	0–1
9	Number	STANDARD == "MA 4.1.1.h" & DOK %in% c(1, 1)	1–2	1
10	Number	STANDARD == "MA 4.1.1.k" & DOK %in% c(1, 2)	0–1	0–1
11	Number	DOMAIN == "NR"	4–6	3–4
12	Number	STANDARD == "MA 4.1.2.b" & DOK %in% c(1, 1)	0–1	0–1
13	Number	STANDARD == "MA 4.1.2.c" & DOK %in% c(1, 1)	0–2	0–1
14	Number	STANDARD == "MA 4.1.2.d" & DOK %in% c(1, 1)	0–1	0–1
15	Number	STANDARD == "MA 4.1.2.f" & DOK %in% c(1, 1)	0–1	0–1
16	Number	STANDARD == "MA 4.1.2.g" & DOK %in% c(1, 1)	0–1	0–1
17	Number	DOMAIN == "NO"	5–6	3–4
18	Number	UDOMAIN == "NUM"	10	7
19	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–10	1–7
20	Number	STANDARD == "MA 4.2.1.a" & DOK %in% c(1, 2)	1–2	0–1
21	Number	DOMAIN == "AR"	1	0–1

22	Number	STANDARD == "MA 4.2.2.a" & DOK %in% c(1, 2)	2–3	1–2
23	Number	DOMAIN == "AP"	2	1–2
24	Number	STANDARD == "MA 4.2.3.a" & DOK %in% c(2, 2)	1–2	0–1
25	Number	STANDARD == "MA 4.2.3.b" & DOK %in% c(2, 2)	1–2	0–2
26	Number	DOMAIN == "AA"	3	1–2
27	Number	UDOMAIN == "ALG"	6	4
28	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–6	1–4
29	Number	STANDARD == "MA 4.3.1.b" & DOK %in% c(1, 2)	0–1	0–1
30	Number	STANDARD == "MA 4.3.1.c" & DOK %in% c(1, 2)	0–1	0–1
31	Number	STANDARD == "MA 4.3.1.d" & DOK %in% c(2, 3)	0–1	0–1
32	Number	STANDARD == "MA 4.3.1.e" & DOK %in% c(1, 2)	0–1	0–1
33	Number	STANDARD == "MA 4.3.1.f" & DOK %in% c(1, 2)	0–1	0–1
34	Number	STANDARD == "MA 4.3.1.g" & DOK %in% c(1, 2)	0–1	0–1
35	Number	STANDARD == "MA 4.3.1.h" & DOK %in% c(1, 2)	0–1	0–1
36	Number	DOMAIN == "GC"	4	2
37	Number	STANDARD == "MA 4.3.3.a" & DOK %in% c(1, 2)	0–1	0–1
38	Number	STANDARD == "MA 4.3.3.c" & DOK %in% c(1, 1)	0–1	0–1
39	Number	DOMAIN == "GM"	1	1
40	Number	UDOMAIN == "GEO"	5	3
41	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–5	1–3
42	Number	STANDARD == "MA 4.4.1.a" & DOK %in% c(2, 2)	0–2	0–2
43	Number	DOMAIN == "DR"	1–2	1–2
44	Number	STANDARD == "MA 4.4.2.a" & DOK %in% c(2, 2)	2–3	0–2
45	Number	DOMAIN == "DA"	2	1–2
46	Number	UDOMAIN == "DTA"	4	2
47	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–4	1–2
48	SUM	POINTS	29–32	20–21

Table A-3
Grade 5 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 5.1.1.a" & DOK %in% c(1, 2)	1–2	0–1
4	Number	STANDARD == "MA 5.1.1.b" & DOK %in% c(1, 2)	0–1	0–1
5	Number	STANDARD == "MA 5.1.1.c" & DOK %in% c(1, 1)	0–1	0–1
6	Number	STANDARD == "MA 5.1.1.d" & DOK %in% c(1, 2)	0–2	0–1
7	Number	STANDARD == "MA 5.1.1.e" & DOK %in% c(1, 1)	0–1	0–1
8	Number	DOMAIN == "NR"	4–6	2–3

9	Number	STANDARD == "MA 5.1.2.a" & DOK %in% c(1, 1)	0–2	0–1
10	Number	STANDARD == "MA 5.1.2.b" & DOK %in% c(1, 1)	0–2	0–1
11	Number	STANDARD == "MA 5.1.2.c" & DOK %in% c(1, 2)	0–1	0–1
12	Number	STANDARD == "MA 5.1.2.d" & DOK %in% c(1, 1)	0–2	0–1
13	Number	STANDARD == "MA 5.1.2.g" & DOK %in% c(1, 1)	1–2	1–2
14	Number	STANDARD == "MA 5.1.2.h" & DOK %in% c(1, 1)	0–2	0–1
15	Number	STANDARD == "MA 5.1.2.j" & DOK %in% c(1, 1)	0–2	0–1
16	Number	DOMAIN == "NO"	5–6	4–5
17	Number	UDOMAIN == "NUM"	10	7
18	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–10	1–7
19	Number	STANDARD == "MA 5.2.1.a" & DOK %in% c(1, 1)	2–3	1
20	Number	DOMAIN == "AR"	2	1
21	Number	STANDARD == "MA 5.2.2.a" & DOK %in% c(1, 2)	3	2
22	Number	DOMAIN == "AP"	3	2
23	Number	STANDARD == "MA 5.2.3.a" & DOK %in% c(2, 3)	1	1
24	Number	DOMAIN == "AA"	1	1
25	Number	UDOMAIN == "ALG"	6	4
26	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–6	1–4
27	Number	STANDARD == "MA 5.3.1.a" & DOK %in% c(1, 1)	0–2	0–2
28	Number	STANDARD == "MA 5.3.1.b" & DOK %in% c(1, 1)	0–2	0–2
29	Number	STANDARD == "MA 5.3.1.c" & DOK %in% c(2, 3)	0–2	0–2
30	Number	DOMAIN == "GC"	3	1
31	Number	STANDARD == "MA 5.3.2.b" & DOK %in% c(1, 1)	1	1
32	Number	DOMAIN == "GO"	1	1
33	Number	STANDARD == "MA 5.3.3.b" & DOK %in% c(1, 2)	0–1	0–1
34	Number	STANDARD == "MA 5.3.3.c" & DOK %in% c(1, 2)	0–1	0–1
35	Number	DOMAIN == "GM"	1	1
36	Number	UDOMAIN == "GEO"	5	3
37	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–5	1–3
38	Number	STANDARD == "MA 5.4.2.a" & DOK %in% c(1, 3)	1–3	1–2
39	Number	STANDARD == "MA 5.4.2.b" & DOK %in% c(2, 3)	1–3	1–2
40	Number	DOMAIN == "DA"	4	2
41	Number	UDOMAIN == "DTA"	4	2
42	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–4	1–2
43	SUM	POINTS	29–32	20–21

Table A-4
Grade 6 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 6.1.1.a" & DOK %in% c(1, 2)	0–2	0–1
4	Number	STANDARD == "MA 6.1.1.b" & DOK %in% c(1, 1)	0–1	0–1
5	Number	STANDARD == "MA 6.1.1.c" & DOK %in% c(1, 2)	0–2	0–1
6	Number	STANDARD == "MA 6.1.1.d" & DOK %in% c(1, 2)	0–1	0–1
7	Number	STANDARD == "MA 6.1.1.g" & DOK %in% c(2, 2)	0–2	0–1
8	Number	STANDARD == "MA 6.1.1.h" & DOK %in% c(1, 2)	0–1	0–1
9	Number	STANDARD == "MA 6.1.1.i" & DOK %in% c(1, 1)	0–1	0–1
10	Number	DOMAIN == "NR"	4–5	3–4
11	Number	STANDARD == "MA 6.1.2.a" & DOK %in% c(1, 1)	0–1	0–1
12	Number	STANDARD == "MA 6.1.2.c" & DOK %in% c(1, 1)	0–1	0–1
13	Number	STANDARD == "MA 6.1.2.d" & DOK %in% c(1, 1)	0–1	0–1
14	Number	STANDARD == "MA 6.1.2.e" & DOK %in% c(2, 2)	0–1	0–1
15	Number	DOMAIN == "NO"	3	1–2
16	Number	UDOMAIN == "NUM"	7	5
17	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–7	1–5
18	Number	STANDARD == "MA 6.2.1.a" & DOK %in% c(1, 2)	1	1
19	Number	DOMAIN == "AR"	1	1
20	Number	STANDARD == "MA 6.2.2.a" & DOK %in% c(1, 1)	0–2	0–1
21	Number	STANDARD == "MA 6.2.2.b" & DOK %in% c(1, 2)	0–1	0–1
22	Number	STANDARD == "MA 6.2.2.c" & DOK %in% c(1, 1)	0–2	0–1
23	Number	STANDARD == "MA 6.2.2.d" & DOK %in% c(1, 2)	0–2	0–1
24	Number	STANDARD == "MA 6.2.2.e" & DOK %in% c(1, 1)	0–2	0–1
25	Number	STANDARD == "MA 6.2.2.f" & DOK %in% c(2, 2)	0–1	0–1
26	Number	STANDARD == "MA 6.2.2.g" & DOK %in% c(1, 2)	0–2	0–1
27	Number	DOMAIN == "AP"	5	3
28	Number	STANDARD == "MA 6.2.3.b" & DOK %in% c(2, 2)	0–2	0–1
29	Number	STANDARD == "MA 6.2.3.c" & DOK %in% c(2, 2)	1–2	0–1
30	Number	STANDARD == "MA 6.2.3.d" & DOK %in% c(2, 2)	0–2	0–1
31	Number	DOMAIN == "AA"	3	1
32	Number	UDOMAIN == "ALG"	9	5
33	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–9	1–5
34	Number	STANDARD == "MA 6.3.1.a" & DOK %in% c(1, 2)	1	0–1
35	Number	DOMAIN == "GC"	1	0–1
36	Number	STANDARD == "MA 6.3.2.a" & DOK %in% c(1, 1)	0–1	0–1
37	Number	STANDARD == "MA 6.3.2.c" & DOK %in% c(1, 2)	0–1	0–1

38	Number	STANDARD == "MA 6.3.2.d" & DOK %in% c(2, 2)	0–1	0–1
39	Number	DOMAIN == "GO"	2–3	1–2
40	Number	STANDARD == "MA 6.3.3.a" & DOK %in% c(2, 2)	1	0–1
41	Number	STANDARD == "MA 6.3.3.b" & DOK %in% c(2, 2)	0–1	0–1
42	Number	STANDARD == "MA 6.3.3.c" & DOK %in% c(2, 2)	0–1	0–1
43	Number	DOMAIN == "GM"	2	1
44	Number	UDOMAIN == "GEO"	5	3
45	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–5	1–3
46	Number	STANDARD == "MA 6.4.2.a" & DOK %in% c(2, 2)	1–3	0–1
47	Number	STANDARD == "MA 6.4.2.b" & DOK %in% c(2, 3)	1–3	0–1
48	Number	STANDARD == "MA 6.4.2.c" & DOK %in% c(1, 2)	1–3	0–1
49	Number	STANDARD == "MA 6.4.2.d" & DOK %in% c(2, 3)	1–2	0–1
50	Number	DOMAIN == "DA"	4	3
51	Number	UDOMAIN == "DTA"	4	3
52	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–4	1–3
53	SUM	POINTS	29–32	20–21

Table A-5
Grade 7 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 7.1.2.a" & DOK %in% c(1, 2)	1–3	0–2
4	Number	STANDARD == "MA 7.1.2.b" & DOK %in% c(1, 2)	0–2	0–1
5	Number	STANDARD == "MA 7.1.2.d" & DOK %in% c(1, 2)	1	0–1
6	Number	STANDARD == "MA 7.1.2.e" & DOK %in% c(2, 2)	1	0–1
7	Number	DOMAIN == "NO"	5	3
8	Number	UDOMAIN == "NUM"	5	3
9	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–5	1–3
10	Number	STANDARD == "MA 7.2.1.a" & DOK %in% c(1, 2)	1–2	0–1
11	Number	STANDARD == "MA 7.2.1.b" & DOK %in% c(2, 2)	0–1	0–1
12	Number	DOMAIN == "AR"	2	1
13	Number	STANDARD == "MA 7.2.2.a" & DOK %in% c(1, 1)	0–1	0–1
14	Number	STANDARD == "MA 7.2.2.b" & DOK %in% c(1, 1)	0–1	0–1
15	Number	STANDARD == "MA 7.2.2.c" & DOK %in% c(1, 2)	0–1	0–1
16	Number	STANDARD == "MA 7.2.2.d" & DOK %in% c(1, 1)	0–1	0–1
17	Number	STANDARD == "MA 7.2.2.e" & DOK %in% c(1, 2)	0–1	0–1
18	Number	DOMAIN == "AP"	3–5	2–3
19	Number	STANDARD == "MA 7.2.3.a" & DOK %in% c(1, 1)	0–1	0–1

20	Number	STANDARD == "MA 7.2.3.b" & DOK %in% c(2, 2)	0–1	0–1
21	Number	STANDARD == "MA 7.2.3.c" & DOK %in% c(2, 2)	0–1	0–1
22	Number	STANDARD == "MA 7.2.3.d" & DOK %in% c(2, 2)	0–1	0–1
23	Number	STANDARD == "MA 7.2.3.e" & DOK %in% c(2, 2)	0–1	0–1
24	Number	STANDARD == "MA 7.2.3.f" & DOK %in% c(2, 2)	0–1	0–1
25	Number	DOMAIN == "AA"	3–5	2–3
26	Number	UDOMAIN == "ALG"	9	6
27	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–9	1–6
28	Number	STANDARD == "MA 7.3.1.a" & DOK %in% c(2, 2)	1	1
29	Number	DOMAIN == "GC"	1	1
30	Number	STANDARD == "MA 7.3.3.a" & DOK %in% c(2, 2)	0–2	0–1
31	Number	STANDARD == "MA 7.3.3.b" & DOK %in% c(2, 2)	0–2	0–1
32	Number	STANDARD == "MA 7.3.3.c" & DOK %in% c(1, 2)	0–2	0–1
33	Number	DOMAIN == "GO"	4	2
34	Number	UDOMAIN == "GEO"	5	3
35	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–5	1–3
36	Number	STANDARD == "MA 7.4.2.a" & DOK %in% c(2, 2)	2	1
37	Number	DOMAIN == "DR"	2	1
38	Number	STANDARD == "MA 7.4.3.b" & DOK %in% c(2, 3)	0–1	0–1
39	Number	STANDARD == "MA 7.4.3.c" & DOK %in% c(2, 3)	0–1	0–1
40	Number	STANDARD == "MA 7.4.3.e" & DOK %in% c(1, 2)	0–1	0–1
41	Number	STANDARD == "MA 7.4.3.f" & DOK %in% c(1, 2, 3)	0–1	0–1
42	Number	STANDARD == "MA 7.4.3.g" & DOK %in% c(2, 3)	0–1	0–1
43	Number	STANDARD == "MA 7.4.3.h" & DOK %in% c(1, 2)	0–1	0–1
44	Number	DOMAIN == "DA"	4	3
45	Number	UDOMAIN == "DTA"	6	4
46	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–6	1–4
47	SUM	POINTS	29–32	20–21

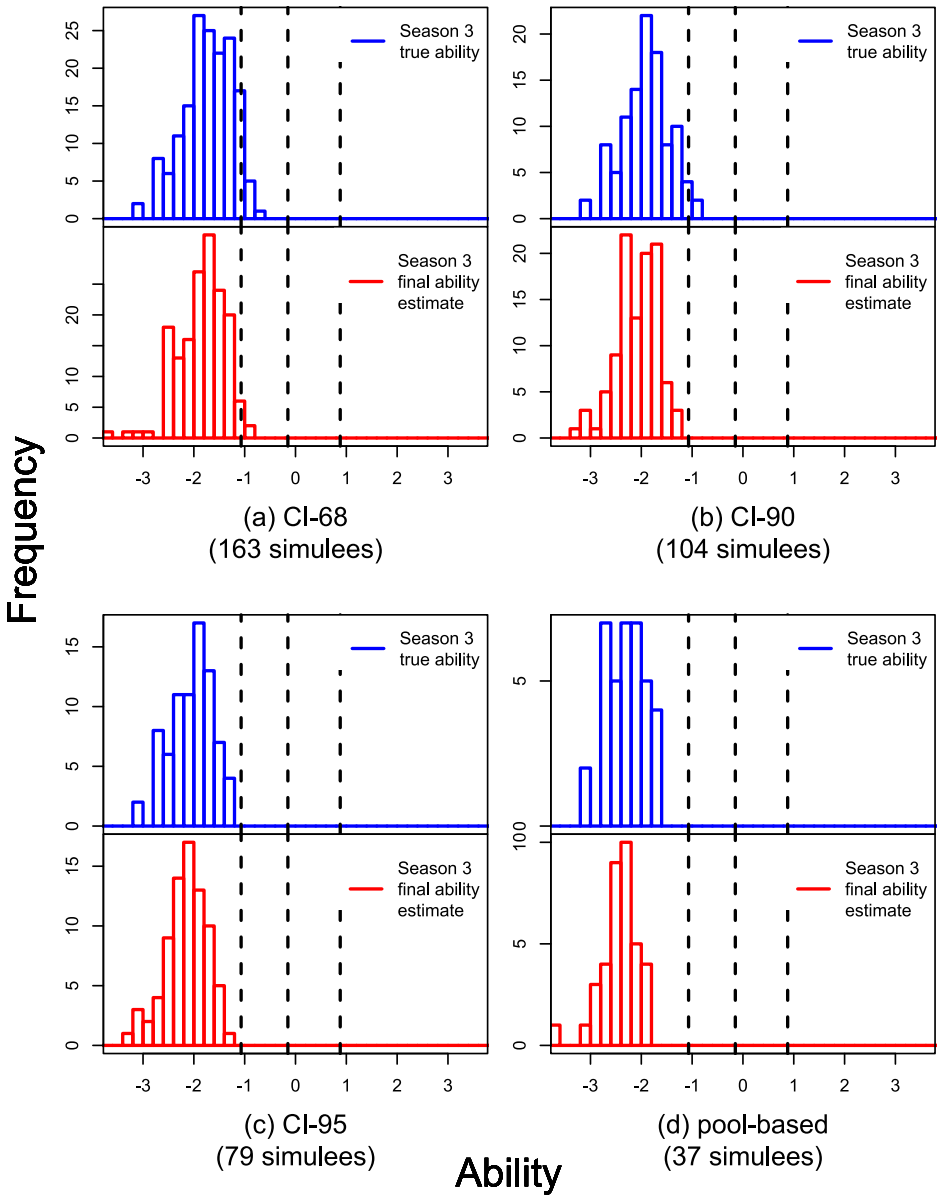
Table A-6
Grade 8 Math Constraints

			Number of Items Range	
Constraint ID	Type	Condition	Phase 1	Phase 2
1	Number		25	16
2	Number	ITEM_TYPE == "Polytomous"	4–7	4–5
3	Number	STANDARD == "MA 8.1.1.a" & DOK %in% c(1, 1)	0–1	0–1
4	Number	STANDARD == "MA 8.1.1.b" & DOK %in% c(1, 1)	0–1	0–1
5	Number	STANDARD == "MA 8.1.1.d" & DOK %in% c(1, 2)	1–2	0–1
6	Number	DOMAIN == "NR"	2–3	1–2
7	Number	STANDARD == "MA 8.1.2.a" & DOK %in% c(1, 1)	0–1	0–1

8	Number	STANDARD == "MA 8.1.2.b" & DOK %in% c(1, 1)	0–1	0–1
9	Number	STANDARD == "MA 8.1.2.c" & DOK %in% c(1, 1)	0–1	0–1
10	Number	STANDARD == "MA 8.1.2.e" & DOK %in% c(1, 2)	0–1	0–1
11	Number	DOMAIN == "NO"	3–4	2–3
12	Number	UDOMAIN == "NUM"	6	4
13	Number	UDOMAIN == "NUM" & ITEM_TYPE == "Polytomous"	1–6	1–4
14	Number	STANDARD == "MA 8.2.1.a" & DOK %in% c(1, 2)	1–2	0–1
15	Number	STANDARD == "MA 8.2.1.b" & DOK %in% c(1, 2)	0–2	0–1
16	Number	STANDARD == "MA 8.2.1.c" & DOK %in% c(1, 1)	0–1	0–1
17	Number	STANDARD == "MA 8.2.1.d" & DOK %in% c(1, 2)	0–1	0–1
18	Number	DOMAIN == "AR"	3–4	2–3
19	Number	STANDARD == "MA 8.2.2.a" & DOK %in% c(1, 1)	1–2	0–1
20	Number	STANDARD == "MA 8.2.2.b" & DOK %in% c(1, 2)	0–1	0–1
21	Number	DOMAIN == "AP"	1–2	1–2
22	Number	STANDARD == "MA 8.2.3.a" & DOK %in% c(1, 1)	0–1	0–1
23	Number	STANDARD == "MA 8.2.3.b" & DOK %in% c(2, 2)	0–1	0–1
24	Number	STANDARD == "MA 8.2.3.c" & DOK %in% c(2, 3)	0–2	0–1
25	Number	DOMAIN == "AA"	3	2–3
26	Number	UDOMAIN == "ALG"	8	5
27	Number	UDOMAIN == "ALG" & ITEM_TYPE == "Polytomous"	1–8	1–5
28	Number	STANDARD == "MA 8.3.1.a" & DOK %in% c(2, 2)	0–2	0–1
29	Number	STANDARD == "MA 8.3.1.b" & DOK %in% c(1, 2)	1–2	0–1
30	Number	DOMAIN == "GC"	1–2	1–2
31	Number	STANDARD == "MA 8.3.2.a" & DOK %in% c(2, 2)	0–1	0–1
32	Number	STANDARD == "MA 8.3.2.b" & DOK %in% c(1, 2)	0–1	0–1
33	Number	STANDARD == "MA 8.3.2.c" & DOK %in% c(1, 2)	0–1	0–1
34	Number	DOMAIN == "GO"	1–2	1–2
35	Number	STANDARD == "MA 8.3.3.b" & DOK %in% c(2, 3)	0–2	0–1
36	Number	STANDARD == "MA 8.3.3.c" & DOK %in% c(1, 2)	0–1	0–1
37	Number	STANDARD == "MA 8.3.3.d" & DOK %in% c(2, 2)	0–2	0–1
38	Number	DOMAIN == "GM"	3–4	2–3
39	Number	UDOMAIN == "GEO"	7	5
40	Number	UDOMAIN == "GEO" & ITEM_TYPE == "Polytomous"	1–7	1–5
41	Number	STANDARD == "MA 8.4.1.a" & DOK %in% c(1, 1)	1–3	1
42	Number	DOMAIN == "DR"	1–3	1
43	Number	STANDARD == "MA 8.4.2.a" & DOK %in% c(2, 2)	1–3	1
44	Number	DOMAIN == "DA"	1–3	1
45	Number	UDOMAIN == "DTA"	4	2
46	Number	UDOMAIN == "DTA" & ITEM_TYPE == "Polytomous"	1–4	1–2
47	SUM	POINTS	29–32	20–21

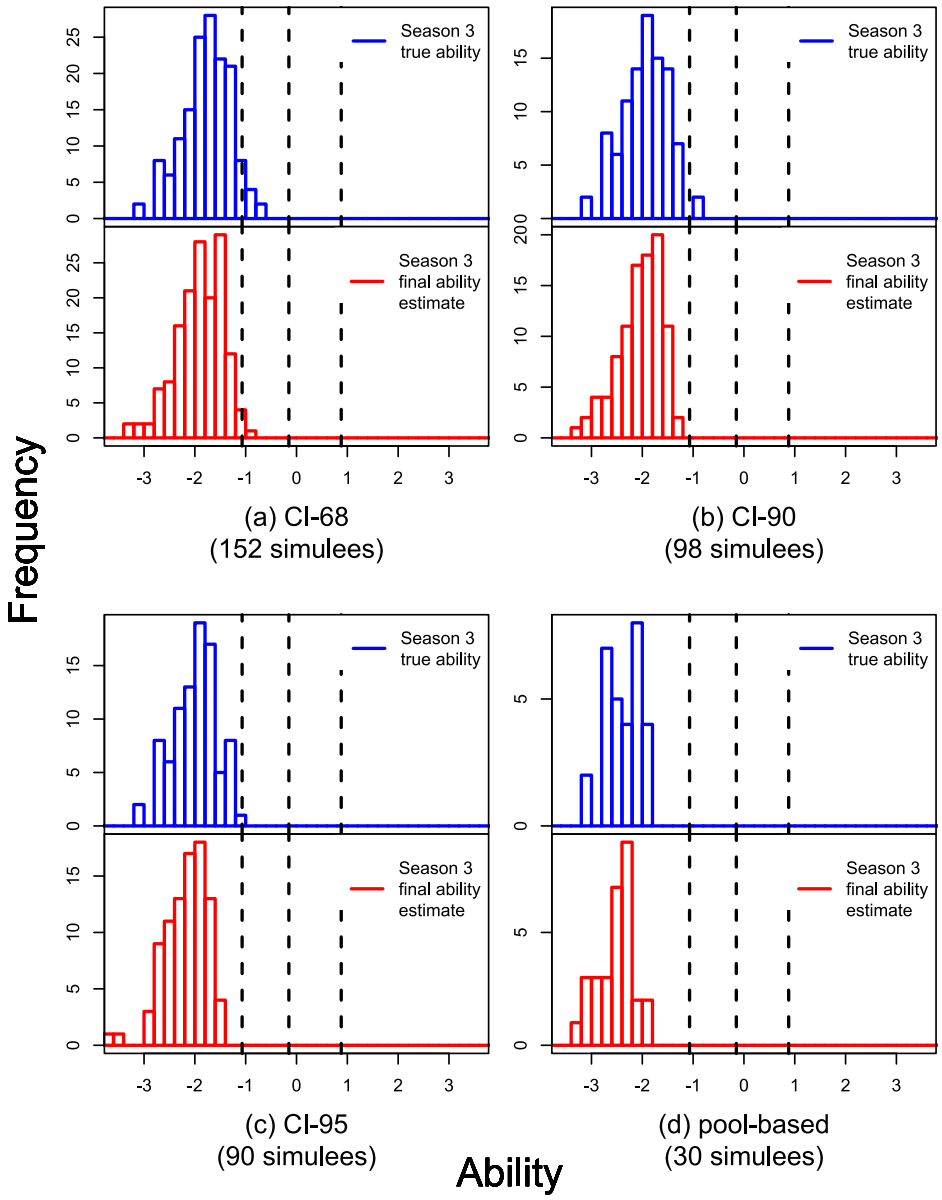
Appendix B: Ability Distributions

Figure B-1
Ability Distributions of Students Routed to Below-Grade in Season 3 (800-Item Banks)



Note. Displayed data are from one replication (four simulations, 1,000 students each) where the grade of record was Grade 4 and item banks had sizes of 800 in each grade. The number of simulees shown under panels (a) to (d) are the number of students who were routed to below-grade (Grade 3) in Season 3 Phase 2. The vertical dotted lines show the cutscores: the first was used for routing below, and the third was used for routing above (the middle cut was not used for routing purposes).

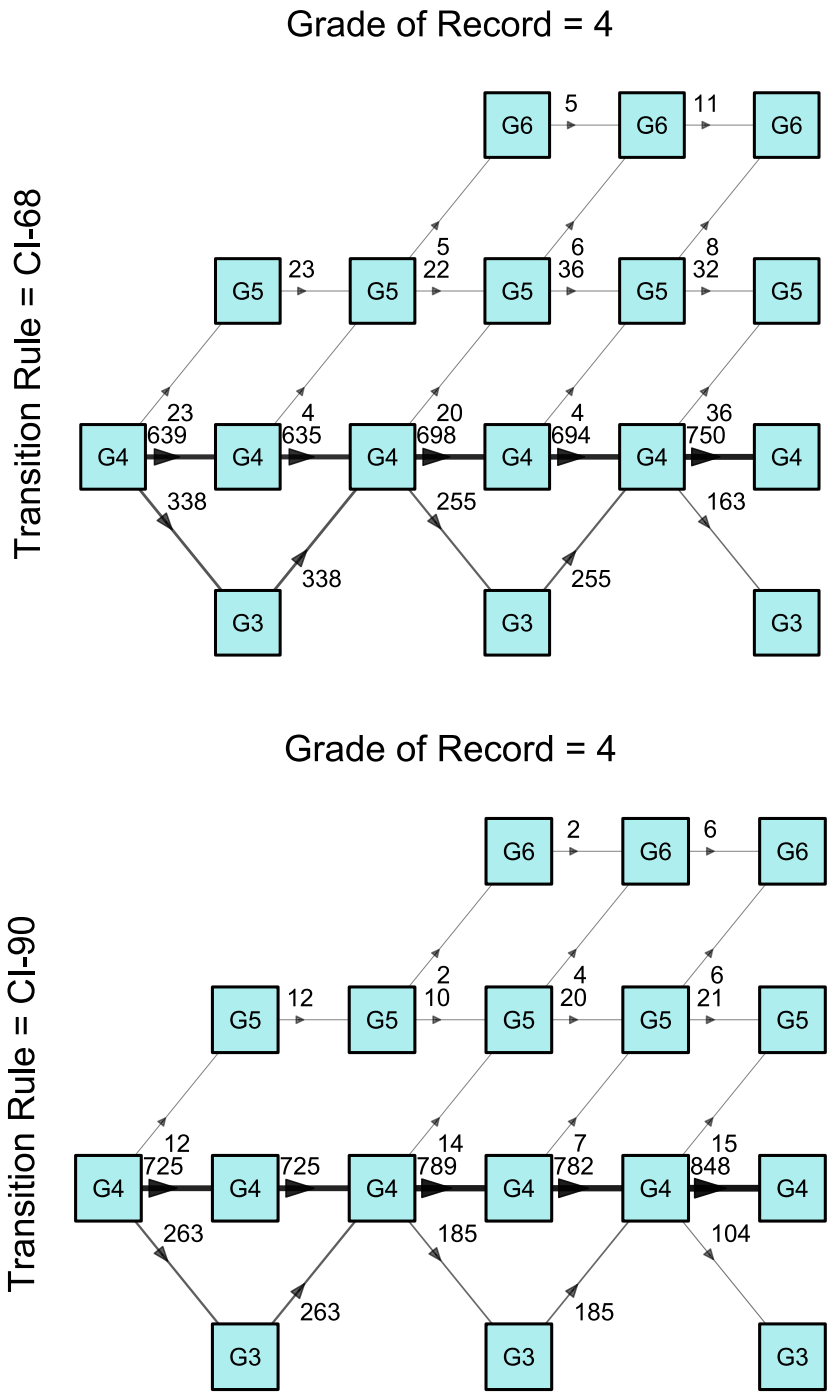
Figure B-2
Ability Distributions of Students Routed to Below-Grade in Season 3 (500-Item Banks)

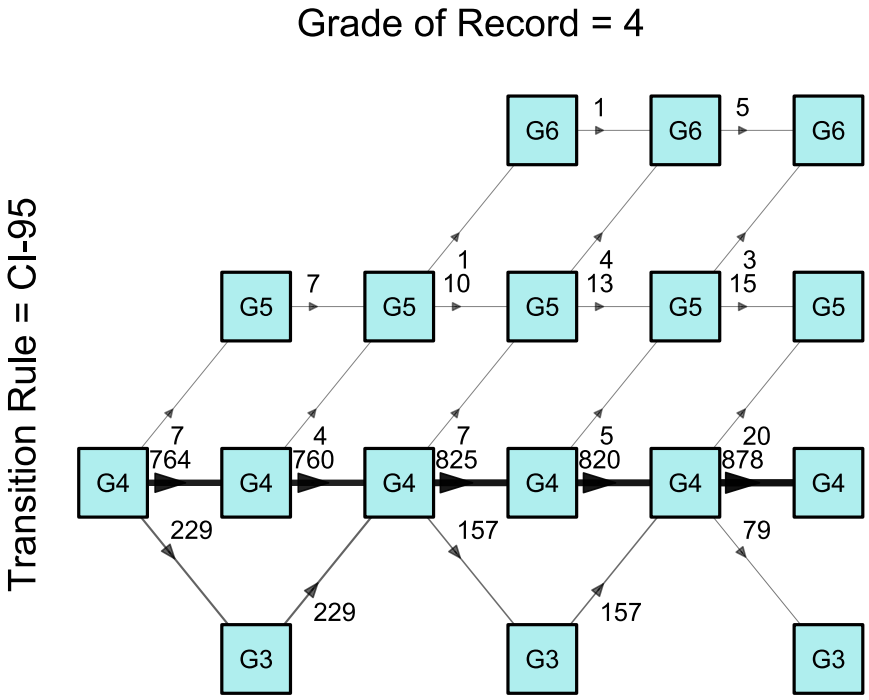


Note. Displayed data are from one replication (four simulations, 1,000 students each) where the grade of record was Grade 4 and item banks had sizes of 500 in each grade. The number of simulees shown under panels (a) to (d) are the number of students who were routed to below-grade (Grade 3) banks and constraints in Season 3 Phase 2. The vertical dotted lines show the cutscores: the first was used for routing below, and the third was used for routing above (the middle cut was not used for routing purposes).

Appendix C

Figure C-1
Number of Students Moving On- and Off-Grade (Grade 4 Simulation)




Grade of Record = 4
Transition Rule = CI-95
G6
G5
G4
G3
1
5
7
1
10
4
13
3
15
7
764
4
760
7
825
5
820
20
878
229
157
79
229
157

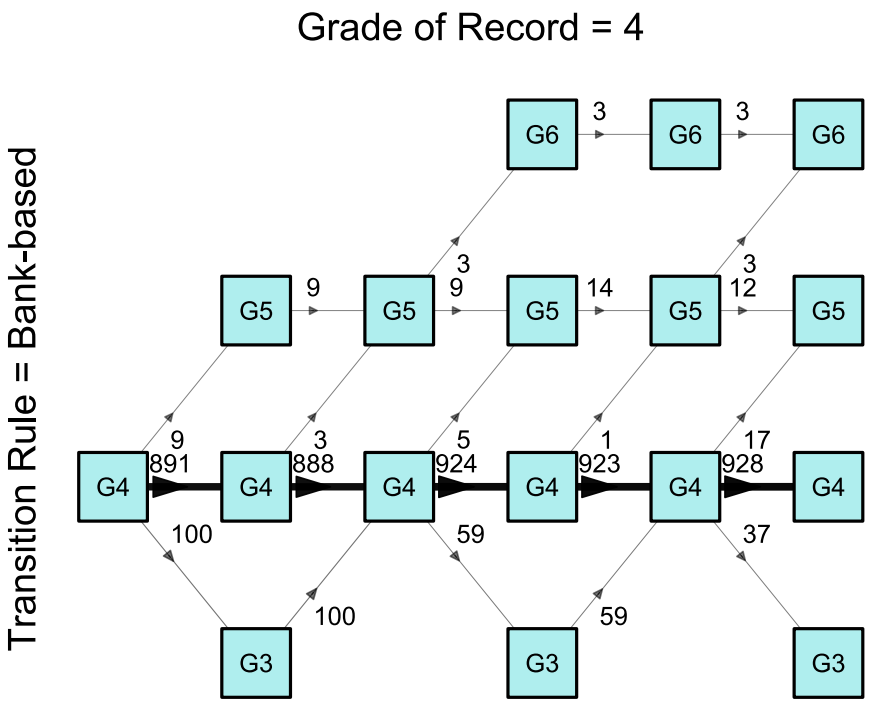
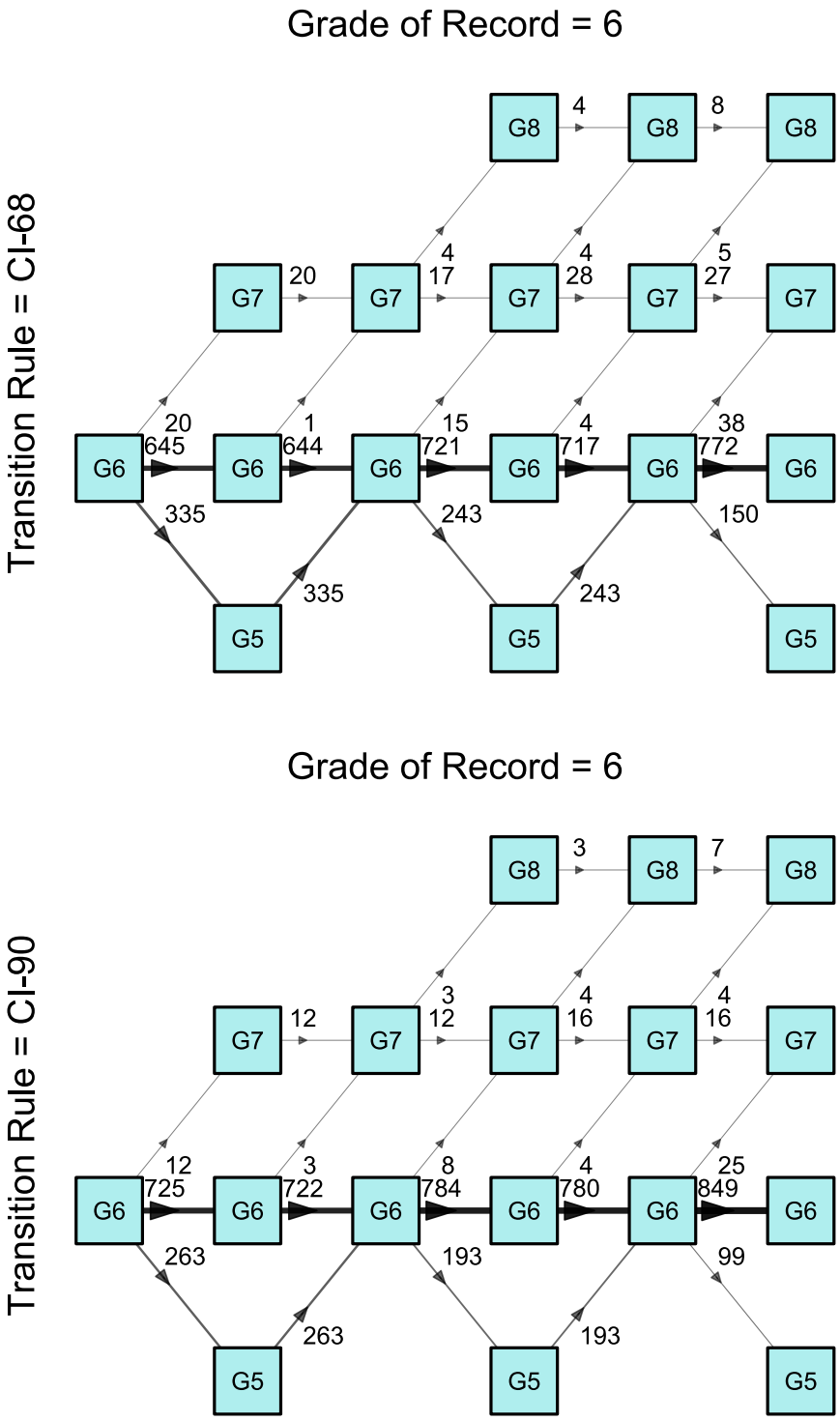
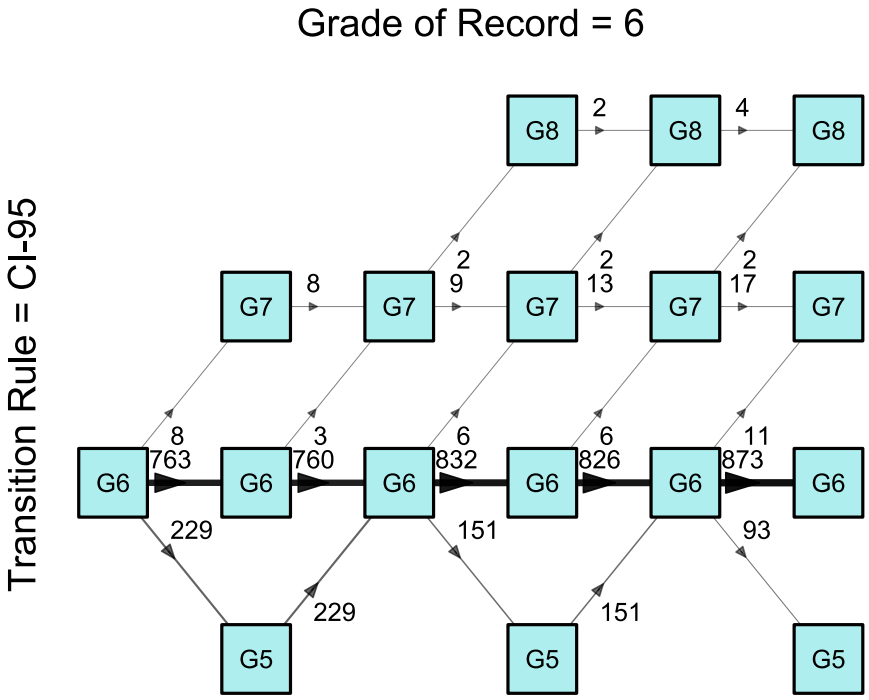
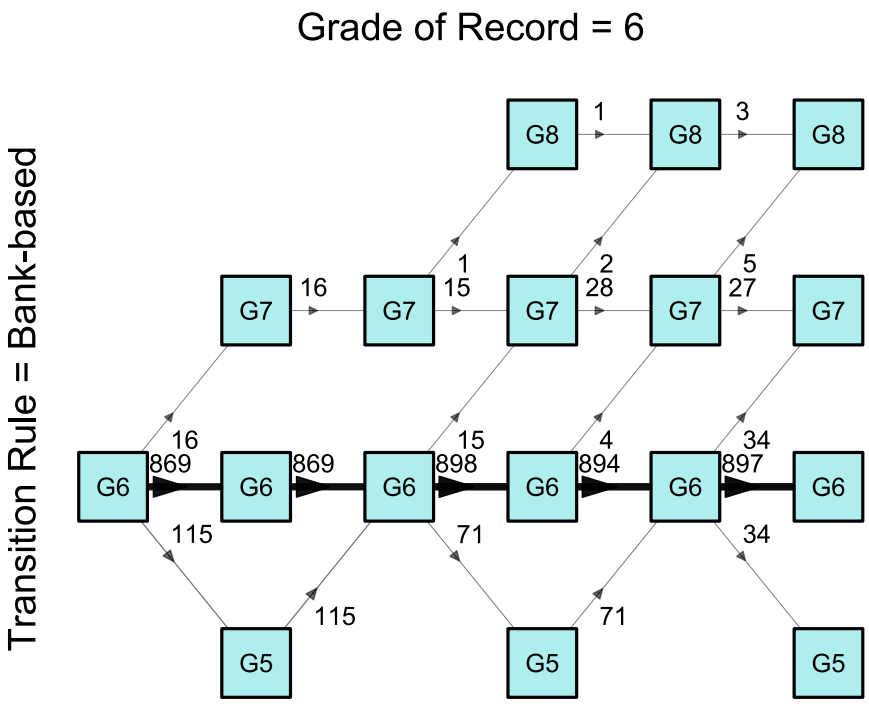

Grade of Record = 4
Transition Rule = Bank-based
G6
G5
G4
G3
3
3
9
3
9
14
3
12
9
891
3
888
5
924
1
923
17
928
100
59
37
100
59

Figure C-2
Number of Students Moving On- and Off-Grade (Grade 6 Simulation)




Grade of Record = 6
Transition Rule = CI-95
G8
G7
G6
G5
2
4
8
2
9
2
13
2
17
8
763
3
760
6
832
6
826
11
873
229
151
93
229
151


Grade of Record = 6
Transition Rule = Bank-based
G8
G7
G6
G5
1
3
16
1
15
2
28
5
27
16
869
869
15
898
4
894
34
897
115
71
34
115
71