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 Amercan 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				
Reporting Category	1	2	3	4	Total Number of Points
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

	Number of Points				
	Curriculum-Embedded PTs				
Reporting Category	Unit 1	Unit 2	Unit 3	End-of-Year Adaptive Test	Total Points
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

	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 Points
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

	Number of Points by End-of-Quarter				
Reporting Category	1	2	3	4	Total Points
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*:

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

Table 10
Blueprint Example: Cumulative Latent Trait Model

	Number of Points												Total Points
	TCSA 1			TCSA 2			TCSA 3			TCSA 4			
Reporting Category	E	I	C	E	I	C	E	I	C	E	I	C	
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*:

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

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):

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

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

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

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 classifycation 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.

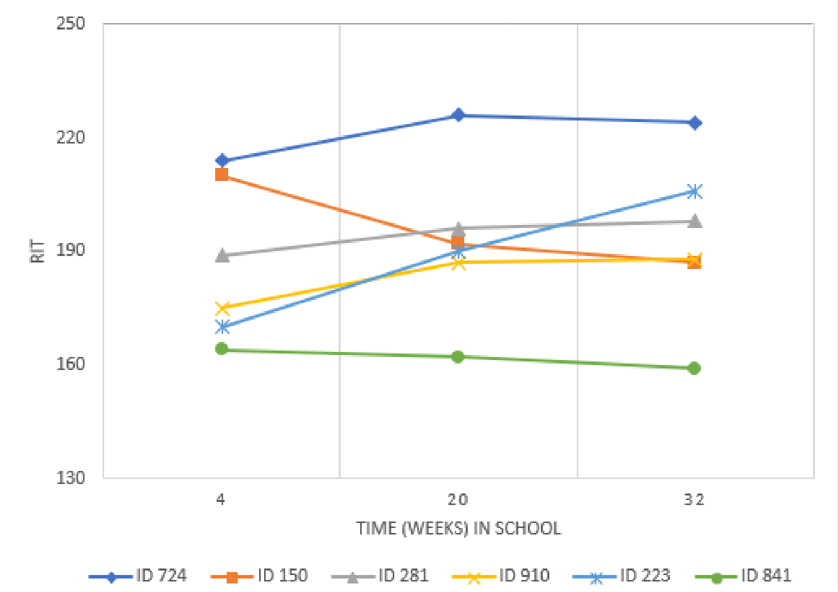
These gaps in the TCSA literature lead to the following research questions that will help guide discussions about an adaptive through-year assessment system:

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