nwea RESEARCH

From Through-Course Summative to Adaptive Through-Year Models for Large-scale Assessment: A Literature Review

July 2019

Garron Gianopulos, Sr. Research Scientist,
NWEA Psychometric Solutions

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Below is a list of abbreviations that appear in this literature review.

ALD	achievement level descriptor
CBAL	Cognitively Based Assessment of, for, and as Learning
CCSS	Common Core State Standards
DAE	distributed accountability exam
DIF	differential item functioning
IRT	item response theory
MIRT	multidimensional item response theory
NAEP	National Assessment of Educational Progress
OTL	opportunity to learn
PT	performance task
RIT	Rasch Unit
TCSA	through-course summative assessment
ATYA	adaptive through-year assessment

1. Introduction

1.1. Literature Review Overview

The purpose of this literature review is to study the advantages and limitations of various through-course summative assessment (TCSA) models with the goal of informing the design of the new and innovative adaptive through-year assessment system at NWEA. This system solution will measure each student's command of grade-level standards and academic growth, while also producing proficiency scores for 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. This model will use a vertical scale unique to each state's academic standards that will link to the Rasch Unit (RIT) scale. This model will be referred to as the adaptive through-year assessment (ATYA) solution to distinguish it from the TCSAs reviewed in this paper. A primary motivator of this literature review is to study TCSAs and identify challenges they pose and to ensure that these challenges are taken into account in the design of the ATYA. Indeed, many of the challenges inherent to TCSAs are not challenges to the ATYA, given its adaptive nature.

This literature review was 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. What are gaps in the literature on TCSA that need further research?

This paper concludes by applying the findings of this literature review to the design of the ATYA, proposing a tentative design of the ATYA model from NWEA that addresses many of the challenges posed by TCSA.

1.2. Definitions

Definitions of key words used throughout this paper are provided in Appendix A. Most are from the *Standards for Educational and Psychological Testing*, hereafter referred to as the *Standards* (AERA, APA, NCME, 2014). An important distinction to be made is between a TCSA and a comprehensive balanced assessment system. While TCSA are most likely derived from comprehensive balanced assessment systems, most comprehensive balanced assessment systems are not TCSAs. TCSA is a newer concept with less appearances in the literature, and the distinction between the two systems is important context for this review.

1.2.1. Comprehensive Balanced Assessment System

A comprehensive balanced assessment system 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 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).

1.2.2. Through-Course Summative Assessment (TCSA) System

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).

> “*A through-course summative assessment system includes multiple assessment components. Components are administered periodically over the course of the school year. Student performance on each component is aggregated to produce summative results*” (Nelhaus, 2010).

The use of the term “aggregate” in Nelhaus’ definition 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 purposes 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* will be defined as a plan that answers the two design questions below. “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 paper vary in how they address these 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?

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, Bennett, & Deane, 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.

1.3. 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 1.1 presents the papers that satisfied this criterion.

Table 1.1. 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

As shown in the table, many papers were written on the topic of TCSA circa 2010, partly in response to the US 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. 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 NCLB era of accountability that had many unintended negative consequences along with their positives.

2. TCSA Model Designs

The literature includes two types of interim blueprint designs: distributed blueprints and cumulative blueprints. 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 takes a representative sample of the summative blueprint at each test event 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: Sum scores, maximum score, simple averages, or weighted averages
2. Complex: Latent trait scales scores or expected scores based on a unidimensional item response theory (IRT) or multidimensional item response theory (MIRT) model

Table 2.1 presents a matrix of seven models found in the literature based on combinations of blueprint designs and aggregation methods.

Table 2.1. TCSA Models based on Score Aggregation Method and Interim Blueprint Design

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

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.

2.1. 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:

- Darling-Hammond and Pecheone's “Balanced Assessment System” (2010)
- Wise’s “End-of-Unit Model” (2011)
- 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?”

2.1.1. Balanced Assessment System

Darling-Hammond and Pecheone's Balanced Assessment System (2010) specifies curriculum-embedded performance tasks (PTs) that measure complex and higher-order thinking skills in interim tests administered after each of three units of instruction. The system gets its name from the balanced use of item types such as PTs, simulations, and multiple-choice items. At the end of the year, a cumulative adaptive test is administered. The PT scores and end-of-year adaptive score are aggregated with weights to produce the summative score (Darling-Hammond & Pecheone, 2010). Table 2.2 illustrates a possible blueprint structure that would support this design. It should be noted that this approach does not eliminate the need for a final summative test event.

Table 2.2. Blueprint Example: Balanced Assessment System

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

*PT= performance task. 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) suggest that the total score could be a weighted combination of PTs 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 PTs would "…comprise from 20–50% of the total score" (Darling-Hammond & Pecheone, 2010, p. 20).

2.1.2. End-of-Unit Model

Wise's End-of-Unit Model (2011) does not specify item types but assumes that the content will be tested after each unit of instruction. This model is designed to "be a better measure of what students knew immediately after instruction in a topic or skill" (Wise, 2011, p. 19). Table 2.3 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 2.3. Blueprint Example: End-of-Unit Model

	#Points				
Reporting Category	**End-of-Quarter 1**	**End-of-Quarter 2**	**End-of-Quarter 3**	**End-of-Quarter 4**	**Total #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, p. 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, it is important that this design is used for content areas in which learning is bounded to each quarter. Finally, it should be pointed out that this model did not specify a linear or adaptive test design.

2.1.3. American Examination System

Resnick and Berger's American Examination System (2010) uses a pretest and posttest design. Each posttest is a distributed accountability exam (DAE) that measures the content taught in the given unit of instruction. During each test event, the student takes 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 the inclusion of 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). Like the previous model, the authors did not specify if the test should be a linear or adaptive test.

Table 2.4 illustrates a possible blueprint structure that would support this design. To keep the example blueprints comparable, all the blueprints in this literature review are 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. Therefore, 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 claim would shorten the length of the DAEs. There is nothing unique about the blueprint design that would prevent this same approach to be applied to any of the TCSA models.

Table 2.4. Blueprint Example: American Examination System

	#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	–	–	**40**
Algebra	5	5	10	10	5	5	**40**
Geometry	–	–	5	5	15	15	**40**
Total	**20**	**20**	**20**	**20**	**20**	**20**	**120**

*DAE = distributed accountability exam. 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 do not provide an exact aggregation model, Resnick and Berger (2010) discuss 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 that they are advocating the use of a weighted combination of posttest scale scores from each DAE using an IRT or MIRT model.

2.1.4. Advantages and Limitations of Distributed Interim Blueprints

Table 2.5 summarizes the advantages and limitations of distributed models

Table 2.5. Advantages and Limitations of Distributed Models

Advantages	Limitations
• High-quality diagnostic feedback relative to other approaches because more testing time can be given to measuring just what was learned since the last interim assessment (Dadey & Gong, 2017). • More instructionally sensitive. • Can produce equivalent scores across districts if the same pacing guide is used. • Summative scores can be easily summed together to arrive at a meaningful total score.	• Breadth of coverage in each interim test may be lost (Dadey & Gong, 2017). • The summative score may not reflect learning loss that has potentially occurred throughout the year. • It does not promote retention (Preston & Moore, 2010). • Requires districts to use common pacing guides or common blueprints. • Should only be used if academic growth is not expected to continue beyond the test event.

2.2. Cumulative Interim Blueprints

A criticism of the distributed blueprint approach is that it does not provide 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:

- Preston and Moore's "Cumulative Balanced Assessment System" (2010)
- Preston and Moore's "Cumulative American Examination System" (2010)

- Wise's "Continuous Learning Model" (2011)
- Zwick and Mislevy's "Cumulative Latent Trait Model" (2011)

2.2.1. Cumulative Balanced Assessment System

To address some of the limitations of distributed models, Preston and Moore (2010) suggested a cumulative version of the Balanced Assessment System. This model is a replica of the original except that each PT is cumulative rather than restricted to just the last unit of instruction.

Table 2.6 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 attain a balance of 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 may not be desirable since the proportion of items should typically match the proportion of instructional time spent on each reporting category.

Table 2.6. Blueprint Example: Cumulative Balanced Assessment System

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

*PT= performance task. Preston and Moore (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.

Preston and Moore (2010) do not provide any details on how the summative score would be produced, but they do state that methodological questions would have to be answered if this approach was used (p. 6).

2.2.2. Cumulative American Examination System

Preston and Moore (2010) also suggested a cumulative version of the American Examination System. This model is a replica of the original American Examination System except that each DAE is cumulative rather than restricted to just the last unit of instruction. Table 2.7 illustrates a possible blueprint structure that would support this design. Similar to 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, even fewer items are available for posttest scores that would presumably form the basis of the aggregated summative score.

Table 2.7. Blueprint Example: Cumulative American Examination System

	#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**

*DAE = distributed accountability exam

Preston and Moore (2010) do 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).

2.2.3. 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 2.8 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.

Table 2.8. Blueprint Example: Continuous Learning Model

	#Points				
Reporting Category	End-of-Quarter 1	End-of-Quarter 2	End-of-Quarter 3	End-of-Quarter 4	Total #Points
Numerical Operations	30	5	5	5	**30**
Algebra 1	–	25	5	5	**30**
Algebra 2	–	–	20	5	**30**
Geometry	–	–	–	15	**30**
Total	**30**	**30**	**30**	**30**	**120**

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 each quarter, emphasizing the importance of learning each quarter’s content equally well. The weighted average would place greater importance on content learned later in the school year, emphasizing the more sophisticated content and retention. The idea behind the use of a maximum score is to give credit to students for their best performance. If students continuously learn through the school year and if the interim test scores are all scaled to maintain scale score equivalence, 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.

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 et al. (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 et al. (2017) compared different aggregation methods with highly correlated interim scores and reported that there was no significant differences between them. This approach of using instructional time as a predictor of score performance is reminiscent of NWEA taking instructional time into account when developing norms for its interim assessment, MAP Growth (Thum & Hauser, 2015, p. 15).

2.2.4. Cumulative Latent Trait Model

Like the Continuous Learning Model, Zwick and Mislevy's approach (2011) assumes that students will accumulate 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):

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

Table 2.9 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.

Table 2.9. Blueprint Example: Cumulative Latent Trait Model

	#Points*												
	TCSA 1			**TCSA 2**			**TCSA 3**			**TCSA 4**			**Total #Points**
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**

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

This model requires the following data:

- A vector of item responses, *x*.
- A vector of curricular variables, *c*, representing the content student *i* was taught.
- A vector of demographic variables, *d.*

The general MIRT model expresses multiple subscores (Θ) as a function of *x*, *c*, and *d*:

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

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 opportunity to learn (OTL). For reporting purposes, it may be most useful to report expected scores on a released test form of items/tasks using Formula 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. Formula 3 will project scores from different forms onto the same set of items/tasks, thereby producing a common metric.

Formula 4 is 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).

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

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

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

Zwick & 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 Formulas 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).

2.2.5. Advantages and Limitations of Cumulative Interim Blueprints

Table 2.10 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).

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.

Table 2.10. Advantages and Limitations of Cumulative Models

Advantages	Limitations
• 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. • 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. • Most likely provides better classification accuracy than a distributed model because the last interim assessment would provide information on the entire domain, making it less vulnerable to deflated or inflated scores from the fall or winter.	• 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. • 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. • Like the distributed models, all but one plan assumed the same pacing guides for all districts and the same blueprint design. • Zwick and Mislevy's model requires OTL surveys.

2.3. Recommendations from the Literature

This section of the paper presents a number of quotes from Wise, Zwick, and Mislevy that are very appropriate to the design of the ATYA. Wise (2011) provided many recommendations at the ETS-sponsored Through-Course Summative Assessment Symposium held in 2010 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, p. 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 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.

“*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*” (Zwick & Mislevy, 2011, p. 27).

“*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. How much data is needed for stable estimation? Are effects for… [curricula differences from districts] 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*” (Zwick & Mislevy, 2011, p. 27–28).

“*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…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, p. 27–28).

3. Expected Advantages, Challenges, and Potential Solutions to TCSAs

3.1. Expected Advantages

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

- Finer-grained feedback due to an increase in the cumulative number of items used in the calculation of summative scores (Preston & Moore, 2010)
- Increased time to score PTs, 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, Kane, & Bridgeman, 2011)
- Increased curricular and assessment coherence because teachers are more likely to see the connections between instruction, standards, and test items (Wilson & Sloane, 2000)
- 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)
- Potentially reduced measurement error because of the increased number of items used for summative scores (Wise, 2011)
- Potentially increased instructional time, assuming that interim TCSAs replace existing interim and summative tests

3.2. Challenges and Potential Solutions

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

- Controlling for OTL may be challenging (Zwick & Mislevy, 2011; Wise, 2011). If a single blueprint is used and different districts follow different pacing guides, some students may 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).
- If the blueprints do not cover cumulative content, the summative score is expected to measure short-term rather than long-term retention (Nelhaus, 2010; Zwick & Mislevy, 2011). If blueprints are cumulative, the tests may take more time than users would like.
- The peer review guidelines may impose test administration requirements that are a burden to districts (Dadey & Gong, 2017; Zwick & Mislevy, 2011).
- Selecting the optimal score aggregation method and blueprint design is challenging because different methods may advantage or disadvantage different growth trajectories (Wise, 2013). Understanding how students grow differently in different content areas and ensuring that the aggregation method matches different growth trajectories may be difficult (Wise, 2013; Bennett et al., 2011).
- Because scores from each interim test would feed into a summative score used for accountability purposes, educators may perceive the tests to be high-stakes, which may generate test anxiety and test preparation activities that reduce instructional time.
- Given Wise's caution to use "proven models for how students learn" 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.

3.2.1. Controlling for OTL

3.2.1.1. Challenge

If a distributed blueprint is used in a TCSA, but districts follow different pacing guides that advocate conflicting instructional sequences, some students may be tested on content they did not have an opportunity to learn. For instance, if a spiraled pacing guide is used in one district, Pythagorean Theorem may be taught continually throughout the school year; in contrast, another district may teach Pythagorean theorem in late winter. In this case, if the distributed blueprint included items that measured Pythagorean theorem in the fall and winter, students in the latter pacing guide would not have had an OTL the measured content in those seasons. 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). 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. Therefore, if a distributed blueprint is used, it is important to ensure that districts reach agreement on a common pacing guide so that the blueprint can be designed to measure the content that was taught in each interim period. However, it may be unrealistic to ask districts to reach agreement on a single pacing guide (Dadey and Gong, 2017). Therefore, the first design challenge of the ATYA is to ensure that students have had the opportunity to learn the content being tested or to ensure that students are not unfairly penalized for not having opportunity to learn tested content.

3.2.1.2. Potential Solutions

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 test engine.

If matching blueprints to different pacing guides is not feasible, the effect of no OTL could be addressed by giving students another chance on the next test event to show knowledge in the area in which they hadn't yet been instructed. For instance, if students did not perform well on questions related to the Pythagorean theorem in the fall test event, then the winter adaptive test could present additional items on Pythagorean theorem. In this approach, before producing the summative score, the fall item scores would be replaced with the winter item scores.

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:

- Remove items that the students did not have an opportunity to learn from the calculation of the total score
- 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 may 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 & Mislevy also discuss "MIRT models that accommodate differential change in item characteristics resulting from different...[opportunities to learn curricula]" (p. 10).

3.2.2. Short-Term vs. Long-Term Retention

3.2.2.1. Challenge

Users of a TCSA summative score may 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.

3.2.2.2. 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. 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.

3.2.3. Peer Review Restrictions

3.2.3.1. Challenge

Dadey and Gong (2017) observed that users of interim assessments like their high degree of flexibility and convenience. However, these features may 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.

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).

3.2.3.2. Potential Solutions

This concern can be addressed with an ATYA developed as an evolution of interim and summative assessment – not as one or the other. The solution could be created intentionally to retain the parts of interim assessment that districts value the most while also meeting state needs and peer review requirements for summative testing. It would be important to educate districts on the differences between the test administration requirements of current interim assessments such as MAP Growth and peer review requirements for summative tests, so they understand what about the interim testing experience will remain the same with ATYA and what will be different. Equally important would be ensuring that states understand the benefits of ATYA over traditional summative tests, including not only supporting districts with timelier grade-level performance data, but also accessing information on student growth from fall to spring for a more complete view of school performance.

3.2.4. Selecting the Optimal Score Aggregation Method and Blueprint Design

3.2.4.1. 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 only measure 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:

- Simple summation (Wise, 2011)
- Maximum score (Wise, 2011)
- Simple averages (Wise, 2011; Dadey & Gong, 2017)
- Weighted averages (Wise, 2011; Ho, 2011; Dadey & Gong, 2017)
- 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 may be time-limited while others may 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.

3.2.4.2. 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 MAP Growth 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?”

3.2.5. Unintended Consequences of a High-Stakes Perception

3.2.5.1. Challenge

Because scores from each interim test would feed into a summative score used for accountability purposes, educators may 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).

3.2.5.2. Potential Solutions

The best antidote to these unintended consequences is a well-designed, balanced assessment system that is comprehensive, continuous, and coherent. To avoid narrowing the curriculum, the item pools must be *comprehensive* so that the full depth and breadth of adopted content standards are measured, which means including a variety of item types that will measure higher-order thinking skills. Large item pools 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 may 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, Thompson, Evans, Martineau, & Dadey, 2018). Shepard, Penuel, and Pellegrino (2018) and Wilson (2018) have argued that learning progressions should act “as the organizing framework for connecting the various assessments and learning activities in a vertically coherent system” (Marion et al., 2018, p. 3). Although there has been considerable optimism around learning progressions, there are also challenges with implementing and validating them.

3.2.6. Building Assessments on Unvalidated Learning Progressions

3.2.6.1. Challenge

Learning progressions are frequently referenced in the 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: “Descriptions of the increasingly more sophisticated ways of reasoning in a content domain that follow one another as a student learns” (Smith, Wiser, Anderson, & Krajcik, 2006). 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 may 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 may 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, Maloney, & Gianopulos, 2017). 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 may 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 may not generalize across school systems.

3.2.6.2. 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 may be more successful in modeling learning progressions. For example, Bayesian networks can be used to connect all the items in the item pool 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 pool 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 achievement level descriptors (ALDs). In fact, ALDs 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 ALDs within a student’s *zone of proximal development,* or “content which the student is ready to learn” (Dupree, 2011, p.1).

With ALDs 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 Range ALDs regardless of whether the teacher uses a classroom, interim, or summative assessment” (Schneider & Veazey, 2018). ALDs are central to test development and score interpretation in a principled test design approach (Schneider & Veazey, 2018) in which “the evidence to draw conclusions is made explicit in the ALDs and items are developed specific to those evidence pieces” (Schneider & Johnson, 2019). While conventional learning progressions may contradict the order of particular pacing guides, micro learning progressions such as ALDs may be more compatible with different pacing guides and can be tested empirically throughout the test development process.

4. Gaps in the Literature

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

- Little empirical, quantitative research has been done on TCSA models.
- All the TCSA models reviewed herein assumed that interim tests were non-adaptive. Therefore, these models may not generalize well to adaptive interim tests, so further research is needed with adaptive TCSAs.
- 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.
- 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.
    - Research should be conducted to test the sensitivity of TCSA scores to different curricula and various within-year growth patterns.
- 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 as NWEA develops an adaptive through-year assessment solution:

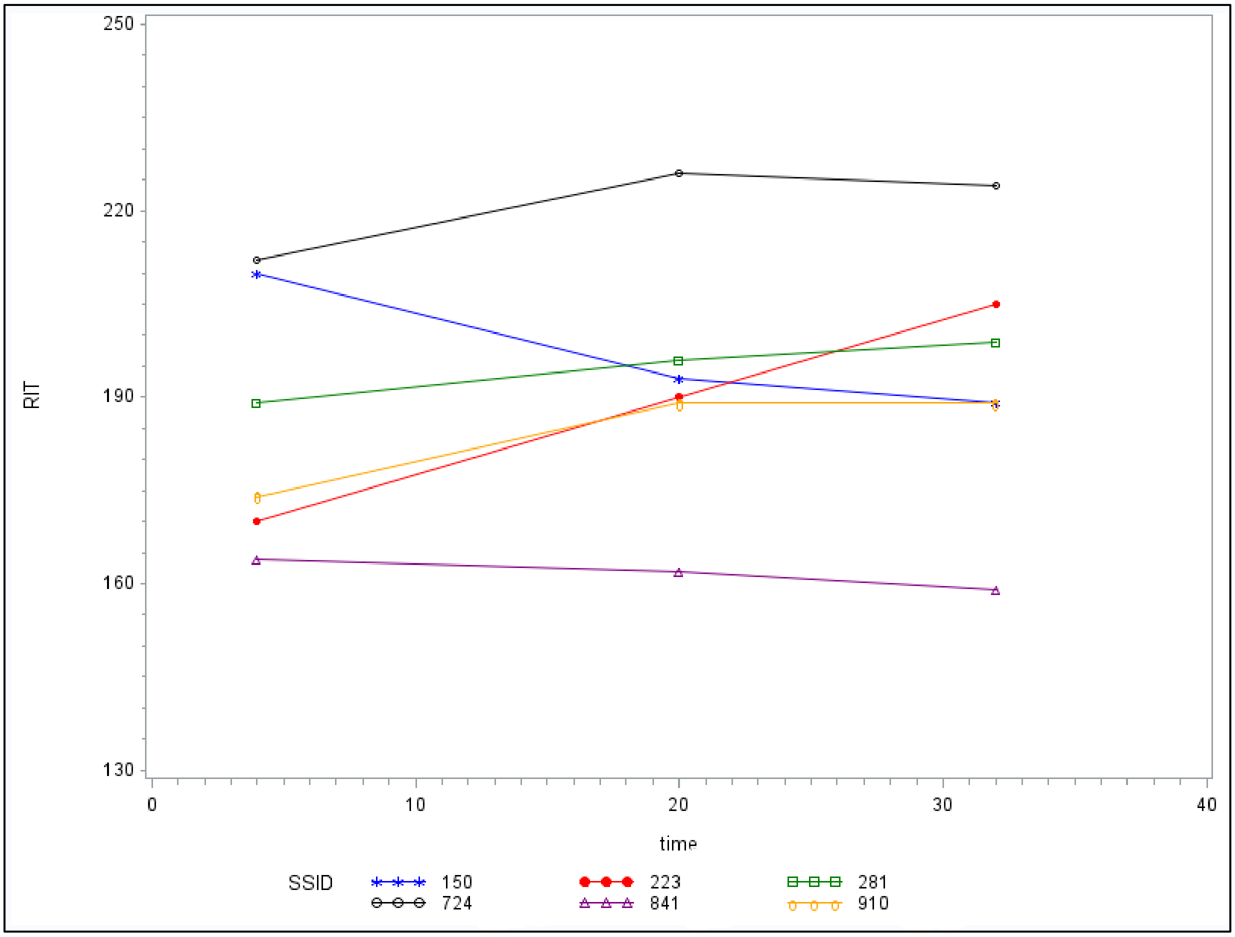
1. How might the use of adaptive interim tests change the advantages and challenges of implementing a TCSA solution?

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, overcome the many challenges of the TCSA approach?

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?
    a. How might a student covariate for curricular variables (*c)* be used to control for OTL in the individual summative scores?
        i. Used for predicting/adjusting IRT difficulty parameters during scoring?
        ii. Used to detect differential item functioning (DIF) between no OTL and OTL during calibration?
        iii. Used as a constraint variable in the adaptive algorithm?

4. 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 the sample of MAP Growth score patterns in Figure 4.1 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 4.1 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 ATYA scoring model provides the best growth measures and the best proficiency classifications? If the ATYA employs a comprehensive blueprint, is it even necessary to aggregate scores to produce a summative score, or can the last test event serve that purpose?

Figure 4.1. MAP Growth Score Patterns



5. Resnick & 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 may 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 may backfire if the prior ability estimate is biased by disengaged test taking as indicated by rapid guessing (Wise, 2017), cheating, or gaming behaviors that might artificially lower early scores (Ho, 2011). Therefore, it is prudent to ask the following: How likely are rapid guessing, cheating, or gaming behaviors? How sensitive is the engine to a biased prior or predicted score? What strategies might mitigate these risks?

6. If an aggregated score is still necessary under the ATYA, 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 proficiency score? This is important because if the spring interim test is comprehensive and provides a better measure of student achievement than a weighted score that uses fall and winter interim scores, the aggregated score would be inferior to simply using the last score. There may 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.

7. What role, essential or not, will learning progressions have in the ATYA? Are learning progressions generalizable enough to work across different pacing guides? How can learning progressions be empirically validated? How can an ATYA be developed and stabilized if it is based on learning progressions that have not yet been validated and are subject to change? How might learning progressions research based on the widespread use of an ATYA yield information about areas where many students tend to make leaps in learning and where many students tend to get "stuck"?

8. How does the best score aggregation method for an ATYA compare to a well-developed comprehensive balanced assessment system that does not require aggregation of scores from across the school year?

5. Conclusion

5.1.1. Literature Review Summary

The purpose of this literature review was to study 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 from NWEA. A significant gap in the literature is a lack of research on interim adaptive tests used for summative purposes. Some of the challenges of TCSAs might be addressed via adaptive tests, but other challenges, such as the aggregation method, remain a thorny problem. Finally, a repeated comprehensive blueprint administered adaptively may obviate the need for an approach that uses aggregated scores. Future research, including intensive Monte Carlo simulation studies and empirical, quantitative studies, should be conducted to answer these questions as part of developing an ATYA design.

An important consideration for the design of the new ATYA solution 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. An advantage to this approach is that each adaptive test event would begin selecting items on the basis of the prior score from the last interim test, improving efficiency of the test. The score from the spring test event would be a valid summative score and should be a measure of what was retained at year's end. Item scores and blueprint coverage from all test events would act as evidence of the student's final standing in the on-grade content.

The new adaptive through-year assessment from NWEA has multiple purposes: 1) to classify students into achievement levels based on state-specific content standards, 2) to measure growth in terms of the state content standards, and 3) to provide RIT scores via an auxiliary scale established through a linking study. How might an adaptive through-year assessment be structured to support these three inferences?

The adaptive test could contain two stages focused on different inferences:

Stage 1: On-grade proficiency
Stage 2: Growth

In stage 1, the items could be constrained to the state's on-grade content standards, but in stage 2, the on-grade constraints may be removed if the student is off-grade. In this approach, if a student is actually on-grade, all the items administered to the student would be on-grade. However, if a student is actually off-grade, they will receive a mixture of on-grade and off-grade items. In this manner, the needed data can be collected to support both proficiency and growth inferences.

References

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

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

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

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

Darling-Hammond, L., & Pecheone, R. (2010, March). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. National Conference on Next-Generation K–12 Assessment Systems. Retrieved from https://www.ets.org/Media/Research/pdf/Darling-HammondPechoneSystemModel.pdf.

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

Ho, A. D. (2011). *Supporting growth interpretations using through-course assessments*. Paper commissioned by the Center for K–12 Assessment and Performance Management at ETS.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2018). *A tricky balance: The challenges and opportunities of balanced systems of assessment*. Center for Assessment. Retrieved from https://www.nciea.org/sites/default/files/inline-files/A%20Tricky%20Balance_092418.pdf.

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

Nelhaus, 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. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/inputmeetings.html.

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). Newark, DE: International Reading Association.

Preston, J., & Moore, J. E. (2010, March). *An introduction to through-course assessment*. Raleigh, NC: North Carolina Department of Public Instruction. Retrieved from http://www.dpi.state.nc.us/docs/intern-research/reports/through-course.pdf.

Resnick, L. B., & Berger, L. (2010, March). *An American examination system.* National Conference on Next-Generation K–12 Assessment Systems. Retrieved from http://www.k12center.org/rsc/pdf/ResnickBergerSystemModel.pdf.

Sabatini, J. P., Bennett, R. E., & Deane, P. (2011, April). *Four years of cognitively based assessment of, for, and as learning (CBAL): Learning about through-course assessment (TCA)*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Sabatini.pdf.

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). Rotterdam, Netherlands: Sense Publishers.

Schneider, C., & Veazey, M. (2018). *Principled test design based on range ALDs.* Portland, OR: NWEA.

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

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.

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.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. NWEA Research Report. Portland, OR: NWEA.

U.S. Department of Education. (2010, April 9). Federal Register Volume 75, Issue 68. Retrieved from https://www.govinfo.gov/content/pkg/FR-2010-04-09/pdf/FR-2010-04-09.pdf.

U.S. Department of Education (2015). *U.S. Department of Education peer review of state assessment systems: Non-regulatory guidance for states*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

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). Rotterdam, Netherlands: Sense Publishers.

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

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

Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.* Center for K–12 Assessment & Performance Management at ETS. Retrieved from https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Wise.pdf.

Wise, S. L. (2017). Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. Educational Measurement: Issues and Practice, 36(4), 52-61.

Wylie, E. C. (2017). *Winsight™ assessment system: Preliminary theory of action* (ETS Research Report No. RR-17-26). Princeton, NJ: Educational Testing Service. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12155.

Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.* Center for K–12 Assessment & Performance Management at ETS. Retrieved from https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Zwick_Mislevy.pdf.

Appendix A: Definitions

The following definitions are from the *Standards for Educational and Psychological Testing* (AERA et al., 2014, p. 215–225) unless otherwise noted.

Accountability system: A system that imposes student performance-based rewards or sanctions on institutions such as schools or school systems or on individuals such as teachers or mental health care providers.

Achievement levels: Descriptions of test takers' levels of competency in a particular area of knowledge or skill, usually defined in terms of categories ordered on a continuum, for example from "basic" to "advanced" or "novice" to "expert". The categories constitute broad ranges for classifying performance.

Assessment: Any systematic method of obtaining information, used to draw inferences about characteristics of people, objects, or programs; a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities, for purpose of drawing inferences; sometimes used synonymously with test.

Composite score: A score that combines several scores according to a specified formula.

Interim assessment: Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals to inform policy-maker or educator decisions at the classroom, school, or district level.

Learning progression: Descriptions of the increasingly more sophisticated ways of reasoning in a content domain that follow one another as a student learns (Smith, Wiser, Anderson, & Krajcik, 2006). They can also describe levels of student thinking (Clements & Sarama, 2014).

Opportunity to learn (OTL): The extent to which test takers have been exposed to the tested constructs through their educational program and/or have had exposure to or experience with the language or the majority culture required to understand the test.

Projection: A method of score linking in which score on one test are used to predict scores on another test for a group of test takers, often using regression methodology.

Score: Any specific number resulting from the assessment of an individual, such as a raw score, a scale score, and estimate of a latent variables, a production count, and absence record, a course grade, or a rating.

Summative assessment: The assessment of a test taker's knowledge and skills typically carried out at the completion of a program of learning, such as the end of an instructional unit.

Validity: The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed.

Weighted scores/scoring: A method of scoring a test in which a different number of points is awarded for a correct (or diagnostically relevant) response for different items. In some cases, the scoring formulas awards differing points for each different response to the same item.