A paper Commissioned by the Accountability Systems and Reporting (ASR)
State Collaborative on Assessment and Student Standards (SCASS)
of the Council of Chief State School Officers (CCSSO)

Using Interim Assessments in Place of Summative Assessments?

Consideration of an ESSA Option

Authored by
Nathan Dadey and Brian Gong
The National Center for the Improvement of Educational Assessment

CCSSO
Council of Chief State School Officers

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Using Interim Assessments in Place of Summative Assessments?
Consideration of an ESSA Option

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

Contents

Acknowledgments:

This report was written in collaboration with the Accountability Systems & Reporting State Collaborative on Assessment and Student Standards under the guidance of Chris Domaleski. We thank Ajit Gopalakrishnan and Brian Laurent for their valuable feedback on an earlier draft.

Suggested Citation:

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.

Introduction

Replacing summative assessment with interim[1] assessment is, for many, a seemingly attractive proposition. The Every Student Succeeds Act (ESSA) allows for such replacement, stating that state's accountability assessments may "be administered through multiple statewide interim assessments" to provide "valid, reliable, and transparent information on student achievement or growth" (ESSA, §1111(b)(2)(B)(viii)). This interest in using multiple interim assessments in place of a single summative assessment is not new. It was also reflected in the Competitive Preference Priority 1 of the now closed Enhanced Assessment Grants Program (U.S. Department of Education, 2016). The grant program called for "approaches to transform traditional, end-of-year summative assessment forms with many items into a series of modular assessment forms, each with fewer items than the end-of-year summative assessment" (p. 5). It also appeared in the Race to the Top grant's course assessment model under which a "student's results from through-course summative assessment must be combined to produce the student's total summative assessment score for that academic year" (Overview information; Race to the Top Fund Assessment Program, 2010, p. 18,178).

The implication of this provision is that the single summative score is to be used as *the* indicator of academic achievement within the state's accountability system (i.e., "academic achievement, as measured by proficiency on the annual assessments" (ESSA, §1111(b)(2)(B)(v)(I))). The single summative score and the system of interim assessments that produce it will therefore need to address the same standards of quality that are addressed by traditional programs of statewide summative assessment. Instead of one single statewide summative assessment, a *system* of multiple interim assessments, collectively, will need to produce a score that is "valid, reliable and transparent" (ibid).

Developing and implementing these kinds of systems of interim assessments 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. There are many technical and practical challenges inherent to such systems, many of which currently lack clear solutions. In addition, any set of assessments used to meet the ESSA interim provision will need to, collectively, meet the requirements of federal-peer review. Current commercially available interim assessments, then, will likely need additional documentation, development, or both. In some cases, commercial interim assessments may not meet a state's needs (e.g., interim assessments designed to measure a specific subset of standards), meaning new interim assessments will need to be developed.

1 Although there is no universally accepted definition of an interim, or benchmark, assessment, here we use the term interim to mean assessments that are (1) administered during instruction and (2) produce results that can be aggregated across students. See Perie, Marion, & Gong (2009) for additional consideration of interim assessments.

The purpose of this document is to consider some critical questions involved in the *initial* design and implementation of such a system:

1. Can stakeholders agree to a common interim assessment system (e.g., the same interim assessments administered statewide)?
2. What content will be covered by each of the statewide interim assessments and what is the relationship of that content to curriculum and instruction?
3. How many interim assessments will be administered? What will the administration windows look like in terms of timing, duration, and order?
4. How will a "single summative score" and related achievement levels be created?
5. Will the separate interim assessment scores, as well as the single summative score, be reliable, precise, and accurate enough for each of the state's purposes and uses?
6. How will the interim assessments be kept secure?
7. How and when will the interim assessments results be collected and reported?
8. Will appropriate accommodations be provided?

This document is organized around these questions. The questions are by no means exhaustive and could easily be restructured in a number of ways. Nevertheless, the questions provide a logical sequence for examining important issues involved with pursuing the ESSA interim provision. These questions are not independent of one another — the way one question is addressed interacts and influences the ways in which the other questions can be addressed. In addition, a negative answer to any one of these questions can cast doubt on a state's ability to adequately implement this ESSA option. For example, stakeholders might come to agreement in multiple areas, but be unable to move forward due to disagreements around the order and timing of assessment administration windows. Also, more stringent requirements in one area may require greater flexibility in other areas. If, for example, the content covered on each interim assessment is aligned to a specific subset of content standards, then the administration windows may have to be flexible.

The ultimate conclusion of this work is that **implementing a system of interim assessments will require a sustained, multiyear effort that goes above and beyond that currently involved in typical summative assessment programs**. This effort involves fostering agreement among stakeholders, then developing, implementing, and maintaining a system of interim assessments based on stakeholder agreement. Many of the issues stakeholders will need to consider are not often addressed in typical summative assessment systems, including the number of interim assessments, the administration windows, and the standards to be covered.

1. Can stakeholders agree to a common interim assessment system?

One starting point is building consensus about the common system of interim assessment. A common interim assessment system is a specific set of interim assessments administered statewide. For example, within each grade the system could consist of a single, off-the-shelf assessment that is administered multiple times per academic year, or the system could consist of several assessments, each of which covers a different group of standards. Agreement on the actual assessments to be given, even if the assessments are in the earliest stages of development, is an important step. Developing agreement will involve considerations of a state's context, in particular, the current assessments administered throughout the state, as well as the conditions and rules necessary for the interim assessment results to be used summatively. Also important is careful consideration of who the stakeholders are and how to involve them.

In states where different interim assessments are used across districts or schools, using one system of interim assessments means that many or all districts and schools will need to adopt a new system of interim assessment. One implicit motivation behind the ESSA provision is that this system of interim assessment can produce the same or similar information produced by currently used interim assessments and also produce a single summative score. A common system of interim assessment, then, would ideally replace other systems. There are costs associated with doing so. Teachers who relied on the old interim assessments will need to adapt to the new interim assessments. Administrators who made longitudinal comparisons based on interim assessment data, potentially across a large number of years, will be temporarily unable to do so.

Implementing the common system of interim assessment while retaining other previously used interim assessments, however, is not without its own set of issues. Districts or schools that implement the common system of interim assessment and retain other interim assessments may run into problems similar to those that sparked the need for assessment inventories. The time needed for testing is one such problem. Increasing the number of interim assessments administered will not only require additional classroom time, but may also complicate school and district schedules. These demands may be seen as unpalatable, particularly if coupled with rigid administration requirements. To avoid or ameliorate these types of problems, when multiple systems of interim assessment are used, the goal should be to augment the common interim assessment system, rather than have multiple systems producing parallel information.

In addition to garnering agreement on the specific assessments making up the system, stakeholders need to understand and support the uses the interim assessment results will be put to, as well as the rules and conditions necessary to support those uses. In terms of the uses of the results, the interim assessment results are intended to be used as the indicator of academic achievement with the state's federally mandated accountability system, as the indicator of proficiency and, potentially, as the indicator of growth. The interim assessment system will, therefore, be subject to similar types of scrutiny as previous statewide summative assessment systems used for accountability purposes.

Agreement that *one* system of interim assessment will be administered and used for the purposes of accountability can be further bolstered by agreement on business rules — rules that underpin the administration of a common system of interim assessment and the use of its results. One broad category of rules deals with student participation. These rules will need to be developed based on the understanding that all eligible students in tested grades and subjects will be required to take a set number of interim assessments. Issues to address include: how participation rates will be calculated to determine whether schools are meeting the 95 percent participation requirements, whether students will be allowed to take make-up assessments, how missing scores will be handled, and the consequences for not meeting participation requirements. Using multiple assessments to produce a single summative score also poses unique challenges to the making of business rules. For example, how will a score be created and counted for a student promoted mid-year? A student might, say, take one interim assessment in third grade at the beginning of the year, be promoted midyear and then take a fourth grade interim assessment at the end of the year. A rule would need to address if, and if so how, a score might be created under these types of circumstances. Another example is how the treatment of missing scores impacts the number of students who are counted as participating. The number of participating students could be much lower if students need to take every interim assessment during the year. Clearly, business rules are not limited to issues of participation alone. The other sections of this document examine a number of other issues germane to business rules. Considering business rules early, and their potential impact, will be valuable in terms of transparency and buy in. Ideally, early agreement on business rules will help prevent differences of opinion later on in the interim assessment system implementation process.

Who the key stakeholders are, and how much agreement is needed to implement a common system of interim assessment, will vary depending on a state's context. District and state administrators and policymakers are key stakeholders, as are principals, teachers, parents, and community members. Agreement at the district-level is particularly important — often decisions about interim assessment programs are made at this level. Input from district level-stakeholders will help identify potential issues early on and also help facilitate transitions from any pre-existing interim assessment systems. Without commitment from districts, a common system of interim assessments is almost certainly bound to fail. However, agreement from district-level administrators and policymakers may not "trickle down" to teachers and other educators.[2]

Stakeholder meetings and working groups are one way to build agreement on implementing a common system of interim assessment. The process of negotiating agreement is likely to involve relaxing some of the constraints under which previous summative assessment systems operated. For example, stakeholders may agree to a common system of interim assessments, contingent on a flexible administration window. Key in these negotiations is recognizing what issues a state can accept compromise on and what the minimum acceptable criteria is for each issue. That is, a state needs to determine how flexible it can be and still use the single summative score within its accountability system.

2 Developing and implementing a state communication plan is one way to help ensure that teachers and other educators know and understand the common system of interim assessment.

2. What content will be covered by each of the statewide interim assessments and what is the relationship of that content to curriculum and instruction?

Agreement on the actual assessments to be given implies that the content (i.e., standards) covered by each assessment is defined. Determining the content of each assessment will likely involve stakeholders considering the advantages and disadvantages of various designs. An important distinction in the way assessment content is structured is whether the multiple interim assessments, administered within the same grade, have the *same blueprint* or *different blueprints*. This section considers these two options, particularly in terms of alignment as well as in terms of curriculum and instruction.

Same Blueprint

When each interim assessment within a given grade has the same blueprint,[3] that blueprint will likely need to cover the content standards in the same fashion that traditional, end-of-year summative assessments do. This type of content coverage is necessary for the interim assessments to align fully to the state content standards. Since each assessment covers the same standards using the same blueprint, the interim assessment results can be placed onto the same scale. If so, then the administration of the multiple interim assessments is essentially a repeated measures design — meaning that changes in student performance can be examined *within* the academic year.

However, covering all of the relevant content standards on each assessment requires that depth be sacrificed for breadth. Unlike assessments designed to align to a limited number of standards, under the same blueprint design the number of items that can be allotted to a given standard or set of standards is necessarily small. In addition, under the same blueprint design students will very likely be assessed on content standards on which they have not yet been instructed. In this case, the interim assessment results could have little utility instructionally, as low scores indicate little beyond the fact that students have not yet been taught the material. Alternatively, such results could be a useful way to determine pre-existing levels of achievement in order to, say, tailor instruction.

Since the blueprints do not vary, the relationships between the assessed content standards, curriculum, and instruction are purely a function of when each assessment is administered. Timing becomes more important when the blueprints differ. Curriculum and instruction vary across schools and districts — meaning that some schools and districts could be disadvantaged if their instructional sequences conflict with the groupings of standards on each interim assessment, with the timing and order in which the assessments are administered, or both. The same blueprint design, therefore, may allow a state to better mitigate concerns around the timing and duration of assessment windows, relative to the differing blueprint design.

3 It is our understanding that, with some notable exceptions, most commercially available interim assessments have the same blueprint across administrations within a grade — that is, within a grade, students are given the same interim assessment multiple times.

Differing Blueprints

Interim assessments that align to a different subset of the state content standards have several potential benefits. One potential benefit is that the differing blueprint design allows a greater number of items to be administered overall, increasing the number of items per standard and resulting in greater depth. This increase in items is, in part, driven by the need for sufficient levels of reliability for scores on each interim assessment. Dividing, for example, a traditional summative assessment of 60 items into 3 interim assessments of 20 items would likely result in low levels of reliability. Thus, each interim assessment would need an increased number of items, translating into a greater number of items in total. A second potential benefit is that sufficient levels of reliability could be reached with fewer items, relative to the same blueprint design, due to the narrowed range of assessed content standards. Another potential benefit is that assessment results on specific subsets of standards may be more relevant to educators than information on the entire set of content standards.

The differing blueprint design also presents challenges. Evaluating alignment is one such challenge. The concept of alignment must be broadened from the consideration of the relationship between a single assessment and a state's content standards, to the relationships between a set of interim assessments and the state's content standards. While current alignment approaches are designed to examine a single assessment, we believe they can be expanded to examine the alignment of a collection of assessments without too much difficultly.

A second challenge arising from the relationships between the assessed content, curriculum, and instruction may prove more difficult to address. A state's content standards represent a consensus about what students should know and be able to do. However, the standards are generally agnostic about *when* during the academic year students should be able to demonstrate such knowledge and abilities. Tying interim assessments to specific subsets of standards requires consideration of when students should demonstrate what they know and can do, or at bare minimum, how standards[4] should be grouped together into specific assessments.

Consequently, the state would likely need to solicit agreement on how the standards should be grouped to create assessment blueprints. One way to do so is to develop a shared understanding of what students should generally know and be able to do at various points during the academic year. Such a shared understanding, however, may be difficult to achieve. Expectations about student learning during the academic year, and the scope and sequence of instruction that supports those expectations, vary from teacher to teacher, school to school, and district to district. Obtaining agreement would likely involve a number of meetings with stakeholders. To be clear, a common or shared curriculum is not a prerequisite to implementing the differing blueprints design. Nonetheless, there does need to be some consensus that allows for assessments to be tied to specific standards.

4 An alternative could be to have multiple sets of interim assessments customized to particular schools and districts. However, this level of customization is likely impossible for most states.

To be feasible, the differing blueprint design may ultimately require greater flexibility at the state-level, relative to the same blueprint design. State-level flexibility could come in a variety of forms. One key type of flexibility is flexibility in administration — a state could allow each school or district to determine the order in which assessments are administered, within very broad assessment windows. Even then the grouping of the content within each assessment may be more similar to some curricular scopes and sequences than others, potentially advantaging some schools and districts. In addition, such groupings of standards could be seen as an implicit endorsement of particular curricular scopes and sequences. Therefore, even if states are flexible in a number of areas such as administration, there is a risk that a system based on the differing blueprint design may be perceived as an effort to homogenize curricular scopes and sequences. Given this potential, stakeholder input becomes even more critical.

3. How many interim assessments will be administered? What will the administration windows look like in terms of timing, duration, and order?

The ESSA interim provision dictates that "multiple interim assessments" (i.e., two or more interim assessments) be used in place of a single summative assessment. Determining the specific number of assessments to be given during the academic year, as well key aspects of the testing windows such as their timing and duration, will likely require consideration of a number of issues, including the state's history of assessment, the design of the interim assessment system, and the logistics of administration.

Assessment History

The ways in which assessments have been administered, and how their results have been used, will likely inform any following assessment system. In considering a system of interim assessments, this history includes both the state's prior summative assessment used for accountability purposes, as well as each district's local assessment initiatives — particularly interim assessments. Scheduling is a key concern. In particular, district assessment schedules are often inundated. Time will need to be found, or made, within the schedules in ways that are amenable to educators and administrators. In addition, scheduling may be difficult if previous practice differs greatly from that proposed under the interim assessment system. For example, the administration of interim assessments is often left up to the discretion of individual teachers. In this case, shifting to a system with specific assessment windows in which all students must participate would require careful positioning by the state to insure adequate support and participation.

Another key concern is the ways in which previous assessments are used. If, say, interim assessments have been used as a progress check in the middle of the year, educators would likely expect any system moving forward to do the same. In some cases, an interim assessment

system will not support all of the uses layered onto to previous interim assessments. And, in some cases, this lack of support will be beneficial, discouraging inappropriate uses of interim assessment scores. In other cases this lack of support may end practices that positively impact student achievement.

Design of the Interim Assessment System

The design of the interim assessment system may determine some or all of the specifics of the administration of the interim assessment system. The differing blueprints design clearly determines the number of interim assessments through the grouping of the content standards into multiple blueprints (e.g., if the standards are grouped into two blueprints, then there will be two interim assessments). The intended use also plays a role in determining the specifics of the assessment windows. For example, the interim assessment results could be used both for accountability purposes and as a part of students' grades — requiring, say, quarterly administration. Under the same blueprint design, the number of required assessments is more directly tied to the intended use of the assessment results. A gain score (i.e., a pretest/posttest difference) requires two administrations. More complex models of change would require three or more administrations.

Logistics

The logistics surrounding the administration of traditional summative assessment often tax local and state educational agencies. The administration of a system of interim assessments is likely to be more taxing, requiring a greater commitment of time and resources. To produce results that can be aggregated into a single summative score, each administration will likely need a number of the kind of logistical supports found in high stakes summative testing programs, including help desks, assessment monitoring, test administration manuals, proctor training, and material distribution. The cost alone may dictate the administration conditions of the interim assessment system.

The amount of time required for testing is another important logistical concern. The combined testing time for the interims will be greater than that of a single summative high-stakes test. This increase is both a function of the number of items students take (although each interim has fewer items than the traditional end of year summative, the total number of items on the combined set of interims will be greater) as well as "fixed" costs of each test administration (e.g., getting students set up and logged onto a computer for computer bases tests, reading directions, collecting and returning paper based tests). The time required for testing itself is often small compared to the amount time needed by educators and administrators to tackle the logistics of test administration. As noted previously, school calendars are generally overflowing and fitting multiple rounds of high stakes testing into these calendars can, and will be, challenging, particularly if the interim assessments are not well integrated with classroom practice (e.g., require pull out testing in a computer lab). A standardized assessment administration that takes an hour could, potentially, require shifts in a school schedule weeks prior to administration.

If the administration of the interim assessment systems is overly prescribed, schools may lose much needed flexibility in their schedule. Flexibility in assessment administration may help alleviate such problems, but can cause others, for example, causing assessment support staff to be "always testing." Given these logistical limitations, an interim assessment system is more likely to succeed if it is (a) able to replace previous programs of both summative and interim assessment, (b) seen as a normal part of classroom practice, instead of as externally imposed, and (c) flexible enough to meet the demands imposed by logistical and practical constraints.

An Aside about Administration Order

Under the different blueprint design, the order in which the interim assessments are administered also matters. Imagine that there are three interim assessments, each aligned to a different subset of the content standards. The first interim assessment is aligned to a subset of content standards, say "A," the second, "B," and the third, "C." Further suppose that the administration of these assessments is fairly rigid — that they are to be administered in order after 60, 120, and 170 days of instruction. However, some districts and schools have curricular sequences following the order B, A, C. These districts may be disadvantaged by the adherence to the order in which the assessments are meant to be administered, particularly if instruction on A does not happen until after the administration of the first assessment.

Increased flexibility in the ordering, as well as the timing, of assessments may help mitigate these types of problems stemming from differences in the scope and sequence of instruction. However, increased flexibility also has the potential to threaten, or at least complicate, the comparability and security of the interim assessment results. Returning to the example above, consider a case in which districts could choose the order in which to administer each assessment within broad administration windows. Suppose schools in one district gave assessments A, B, and C after 60, 120, and 170 days of instruction, whereas schools in another district gave assessments C, A, and B after 20, 80, and 140 days of instruction. Are the single summative scores comparable? Could the items on assessment A be exposed to students in the second district? Moreover, would the accountability actions attached to the single summative assessment scores be perceived as fair? Currently, answers and approaches to these questions are in short supply. These issues would likely need to be the consideration of special studies, studies built into the ongoing cycles of assessment maintenance.

4. How will a "single summative score" and related achievement levels be created?

For the purposes of accountability, ESSA requires a "single summative score that provides valid, reliable, and transparent information on student achievement or growth" (ESSA, §1111(b)(2)(B)(viii)). This single summative score is not defined further, nor is it mentioned anywhere else within the law. The implication is that the single summative score will need to satisfy

the same criteria as scores from traditional summative assessments. Given prior practice on the reporting of assessment scores, as well as the final ESSA regulations on Accountability and State Plans (2016), this summative assessment score will need to be a proficiency or achievement level classification — in other words, a single summative proficiency level.[5] In addition, states may want to also produce a scale score, a single summative scale score.

In the following sections, we examine the creation of these two types of single summative scores under the two scenarios defined in the prior section — when the interim assessments have the same blueprint, and when they are different. Here, these designs signal whether the interim assessments within a given grade can be placed on the same scale. Under the same blueprint design, the interim assessments can readily be placed onto the same scale, whereas under the different blueprint design, they cannot. Each design supports specific definitions of the single summative score better than others. (In particular, the same blueprint design does not support growth measures well.)

An Aside about Item Data

In our considerations below, we generally assume that the single summative score is produced by combining test scores from each interim assessment — that is, each student receives a single score on each interim he or she takes. The question, then, is how each student's multiple scores should be combined. An alternative approach is to use student item response data directly. Our view is that the key issues are similar, whether item or test-level data are used.

Same Blueprint

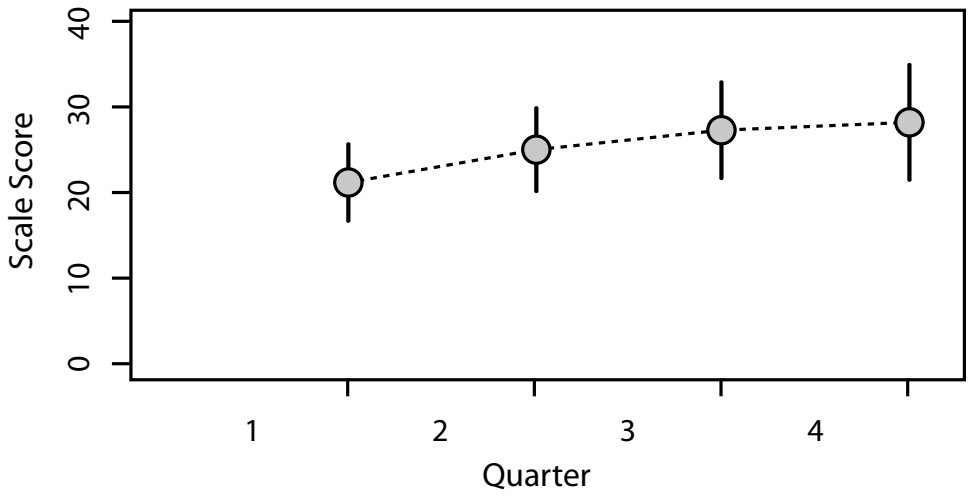
Under this blueprint design, students are assessed on the same set of standards multiple times a year. Students take the same interim assessment multiple times over the course of the academic year, or they take assessments with the same or similar blueprint, but different items. Much of current practice aligns well to this design. Districts that give interims more than once a year often give the same interim in each administration. This design allows assessment results to be placed on the same scale, facilitating the creation of growth measures. However, this design poses conceptual challenges for creating a single status measure.

Scale scores

Generally, the expectation is that student performance improves during the course of the academic year. If performance is changing, how might it be summarized in a single summative scale score? Consider the figure below, which shows a hypothetical distribution of student interim assessment scale scores. Assume that students took the same interim assessment at the end of each quarter of instruction and had scores that increased on average.

5 Although we focus on proficiency level classifications, our reasoning could be easily extended to achievement level classifications.

Figure 1. Hypothetical Distribution of Interim Assessment Scale Scores by Quarter



Note: The grey points represent means, while the vertical lines represent +/- one standard deviation.

What score can be created to summarize changing student performance? One answer might be that a single summative scale score is not enough and that scores corresponding to both initial status and growth are needed. However, our understanding of ESSA is that that the indicator of academic achievement, which the single summative score is meant to replace, is a status measure (ESSA, §1111(c)(4)).[6]

How might a status measure be derived based on results like those shown in Figure 1? Determining how to create a single summative scale score largely depends on what statement or claim a state wants to make about its students' performance. Should the claim be about average student performance in a fashion similar to the way course grades or grade point averages are defined? Should the claim be about a student's best performance, similar to the way a student's best work is selected for a portfolio? Or should the claim be about a composite that weights each assessment according to some value judgment, similar to the way different kinds of work contribute more or less to a student's course grade?

Each of these claims and their related scores could be justified under the ESSA interim option, as could numerous approaches not articulated above. Providing a justification that articulates why a specific approach was taken is key. Such justification could take into account input from stakeholders, as well as considerations of the prior assessment system and psychometric considerations like measurement error. Wise (2011) conducted simulations that provide insight into these last two considerations. Specifically, he simulated scores from four quarterly interim assessments and compared the results to "true" end of the year performance, as defined by his model. He found that average scale scores tended to underestimate end of the year performance, that maximum scale scores tended to overestimate end of the year

6 The only exception is in high school where the indicator of academic achievement can combine status and growth measures.

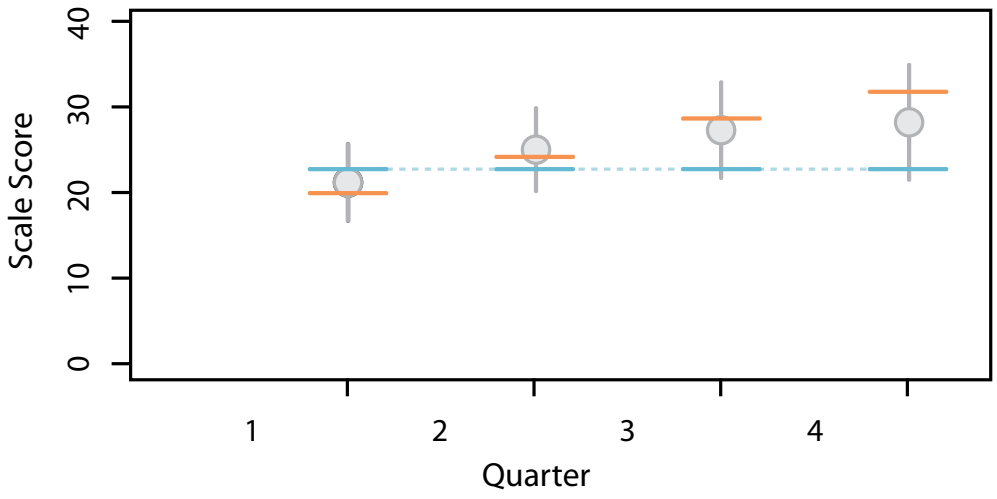
performance, and that a weighted composite score[7] tended to accurately estimate end of the year performance.

Achievement level classifications

The regulations for ESSA (U.S. Department of Education, 2016) stipulate that the indicator of academic achievement be defined based on the percent of students whose academic achievement is at, or above, proficiency. There are at least two approaches that can be taken to produce proficiency level classifications.

One approach is to create classifications for each student at each assessment administration, then aggregate across these classifications to produce the single summative achievement level. The rule used to aggregate the results could be simple numeric operations like taking the median, or could be based on value judgments like those involved in creating a value table. Within this approach, an important question is whether the cut scores used to create the classifications are the same across assessment administrations. Consider a revised version of Figure 1 below. One set of hypothetical cut scores, shown in blue, is the same across administrations. Another set of cut scores, shown in orange, increases across assessments. Clearly, each approach leads to different classifications of students at each time point and ultimately different summative achievement levels. A second approach depends on the creation of a single summative scale score. Once created, a proficiency cut can be applied directly to that score.

Figure 2. Hypothetical Distribution of Interim Assessment Scale Scores and Proficiency Cuts by Quarter



Both approaches require a cut score. If a state is using a pre-existing assessment, it may be tempting to use that assessment's previously set cut score to define proficiency. However, relying on such cut scores may be problematic. Under each approach, the distribution of scores in question is likely to be different than the one used to create the previous cut scores. Instead of

7 This composite was created by weighting the four scores as a function of the cumulative amount of instructional time students would have received — 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.

using pre-existing cuts, therefore, a new standards-setting process should be conducted using data from the individual assessment administrations or the single summative scale score.

A question of growth

The same blueprint design clearly lends itself to the creation of "within year" growth scores. One can easily imagine creating any number of growth scores given patterns like those in Figure 1. Such growth scores could be used solely for purposes of instruction. Alternatively, they could be used as part of a state's accountability system. These within-year growth scores could be used as one of the indicators allowable under ESSA, as a separate indicator in elementary and secondary schools, and as part of a combined status and growth indicator in high school (ESSA, §1111(c)(4)(B)).

Differing Blueprints

When the blueprints of the interim assessment differ, the results must be combined not only to satisfy ESSA, but also to represent the full set of content standards. Potentially, the results of each assessment could be treated as "if they were different sections of the same test" (Wise, 2011, p. 11) and thus combined through psychometric scaling (e.g., subjecting all of the item response data to one item response theory calibration) or through the aggregation of each assessment score into a single composite index or scale (e.g., adding the scores from each interim together, as is often done with assessment batteries).

Scale scores

Assuming a score is produced for each interim assessment, approaches to creating a single summative score are similar to those addressed under the same blueprint design. These approaches, though, do not run into the same conceptual problem posed to approaches under the same blueprint design. Namely, these approaches do not need to directly summarize changing student performance. Instead, the conceptual problem is determining how to combine the assessment scores to most appropriately represent the content standards.

Appropriate representation of the content standards could be achieved directly through the design of the blueprints or through the weighting of items associated with particular standards. For example, a standard might be included on every interim assessment to provide feedback to teachers, but doing so results in the standard being overrepresented. Thus, items aligned to those standards might be weighted less in the creation of each interim assessment score, and therefore in the single summative scale score. Determining whether the collection of interim assessment blueprints appropriately covers the standards requires a clear conception of the depth and breadth of standards to be assessed. That is, there needs to be an overarching idea of how the standards should be represented across the multiple interim assessments. One way to do so is by developing a blueprint that spans all of the interim assessments — a "master" blueprint.

We have conducted an initial small scale[8] study into the aggregation of interim assessments with differing blueprints using empirical data. Although we did not explore weighting assessment results according to a master blueprint, our findings may provide insight into the impact different approaches to weighting assessment results may have. Our findings show that three aggregated scores — mean scale scores, maximum scale scores, and weighted scale scores — correlated so highly (an average Pearson correlation of 0.96) that the method of aggregation matters little. These associations are so strong that other aggregation approaches we did not consider are very likely to come to the same conclusions. In addition, the three scores also had similarly strong associations (an average Pearson correlation of 0.82) with scores from a traditional, end-of-year summative assessment. These *preliminary* findings suggest that students who perform well on one interim assessment perform well on another, and vice versa. In retrospect, this result is not surprising. However, its implication — that the aggregation approaches used to create a single summative scale score perform similarly — is.

Achievement level classifications

The two general approaches outlined for the same blueprint design (aggregating the proficiency classifications for each interim assessment or creating classifications based on a single summative scale score) apply to the differing blueprint design as well. However, if classifications are created for each assessment, they would have to be unique to each assessment, by necessity.

5. Will the separate interim assessment scores, as well as the single summative score, be reliable, precise, and accurate enough for the state's purposes and uses?

Each interim assessment will likely need to produce a score sufficiently reliable for use in local decisions, and the single summative score will need to have levels of reliability for use in high stakes decisions (e.g., as an indicator in an ESSA compliant accountability system). When the interim assessments can be conceptualized as different sections of the same test (i.e., under the differing blueprint approach), the reliability can be readily computed using common methods, like the formulas for "composite" scores (e.g., Haertel, 2006, p. 76-78). In addition, each interim assessment under this design assesses a narrower subset of content standards, meaning that each assessment may be as reliable as a traditional summative assessment with fewer items.

When the interim assessments have the same blueprint, the applicability of formulas for composite scores is unclear. We are not aware of research investigating the reliability of composite scores when they are defined at different points in time. Thus, the reliability of a single summative score based on interims with the same blueprint remains an open question that requires further research.

8 We examined interim assessment results in 6th grade mathematics from a large school district. The items on the 6th grade interim assessments were generally aligned to the standards taught in the quarter it is administered, according to the district's pacing guide. Approximately 5,000 students took interim assessments at the end of the first three quarters and then a traditional end-of-year summative assessment at the end of the fourth quarter.

6. How will the interim assessments be kept secure?

The usefulness of interim assessments often depends on their relative lack of security — that interim assessments can be administered at any time at an individual teacher's discretion without support or monitoring. This flexibility in administration, coupled with the immediate or near immediate reporting of results (for completely electronically scored assessments), allows teachers to make instructional decisions in near real time.

Increasing the security of the assessment administration process will likely constrain the usefulness of the interim assessments. The question is, "Can the interim assessments be made secure enough to be used for high stakes purposes, but also retain features that make it useful to educators?"Peer review guidance (U.S. Department of Education, 2015) requires that assessment programs have a variety of procedures in place to ensure that assessments and their results remain secure, including "proper test preparation guidelines and administration procedures, incident-reporting procedures, consequences for confirmed violations of test security, and requirements for annual training at the district and school levels for all individuals involved in test administration" (p. 30). In addition, Peer Review asks for procedures that identify and address breaches of test security. Implementing all of these procedures will require work and shifts in the ways teachers and other educators think about and administer interim assessments. Addressing the issue of security is particularly daunting because of the number of times students may be exposed to the same or similar assessment. Flexibility in administration, while a solution to numerous concerns like those around curriculum, can also pose serious challenges to security. For example, the window in which assessment content could be inappropriately accessed can be much greater under the interim assessment system, relative to prior systems. This window could be on the order of months, if, say, one district administers an interim assessment at the beginning of the year and another district does so many months later. Finally, to minimize exposure, items may need to be replaced with greater frequency than they would have been under prior assessment programs.

7. How and when will the interim assessments results be collected and reported?

Typically, interim assessment results — if they are collected at all by educational agencies — are collected at the district-level. A state that uses a system of interim assessments would need to negotiate access to interim assessment results. This type of access may require additional work to link district and state databases together.

In addition, interim assessments often provide instant feedback when administered via computer, whereas the results of traditional summative assessments are provided only after equating, human scoring, and other procedures along with related quality control processes, are implemented. Generally, traditional summative assessment results are delivered months

after administration. Whether interim assessment results, that are also used to produce a single summative score, can be returned in time to meet stakeholder needs is another open question.

8. Will appropriate accommodations be provided?

Traditional summative assessments are generally accompanied by a number of accessibility features and accommodations that remove construct-irrelevant barriers for students with disabilities or limited English proficiency. Generally, there are fewer accessibility features and accommodations available for interim assessments than for the traditional summative assessments. To be fair and equitable, as well as to pass federal peer review, interim assessments need to have accessibility features and accommodations similar to those in place for traditional, end-of-year summative assessments. In addition, ESSA regulations also seem to imply that states need to provide translated versions of their general assessments for the most populous language other than English (Inclusion of all students, 2016). Addressing these requirements will take a substantial investment in interim assessment development and implementation.

The use of an interim assessment system also poses questions regarding the state's alternate assessment based on alternate achievement standards. Given limited resources, a state transitioning to an interim assessment system will likely do so first for the general population. Is it advisable for a state then to have an interim assessment system for the general population, but a traditional end-of-year summative assessment for students with the most significant cognitive disabilities? Can an interim assessment system approach be developed and implemented for students with the most significant cognitive disabilities? Like previous questions, addressing these questions, as well as other questions regarding the assessment of students with the most significant cognitive disabilities within an interim assessment system, will require further research.

Discussion

This work addresses some important initial questions that arise from the ESSA interim assessment option. 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. We recommend that interested states investigate this option over several years, starting with small scale investigations and gradually building capacity and knowledge.

Finally, it is important to note that the challenges highlighted in this document are not insurmountable. With diligent planning and sufficient resources, a state may develop a system of interim assessments that is less divorced from day-to-day instruction than traditional summative assessments, and, hopefully, more accepted.

References

Elementary and Secondary Education Act of 1965, as Amended by the Every Student Succeeds Act-Accountability and State Plans, 81 Fed. Reg. 86076, 86076-86248 (November, 29, 2016) (to be codified at 34 CFR pts. 200 and 299).

Every Student Succeeds Act §1111, S.1177. (December 10, 2015). 114th Congress. Retrieved March 13, 2017, from https://www.congress.gov/bill/114th-congress/senate-bill/1177/text.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.) (65-110). Westport, CT: Praeger Publishers.Inclusion of all students, 34 CRF § 200.6. (2016). Retrieved March 14, 2017, from http://www.ecfr.gov/cgi-bin/text-idx?SID=afc4098de740e1b4999297f7ea5042f0&mc=true&node=se34.1.200_16&rgn=div8.

Overview information; Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. 75 Fed. Reg., 18171, 18171-18185 (April 9, 2010).

Perie, M., Marion, M., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13.

U.S. Department of Education. (2016). *Application for new grants under the Enhanced Assessment Instruments Grant Program (EAG)* (CFDA 84.368A). Retrieved December 15, 2016, from http://www2.ed.gov/programs/eag/eag2016application.pdf.

U.S. Department of Education. (2015, September). *Peer review of state assessment systems: Non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended*. Retrieved December 15, 2016, from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.Wise, L. (2011). *Picking up the pieces: Aggregating results from through-course assessments*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072