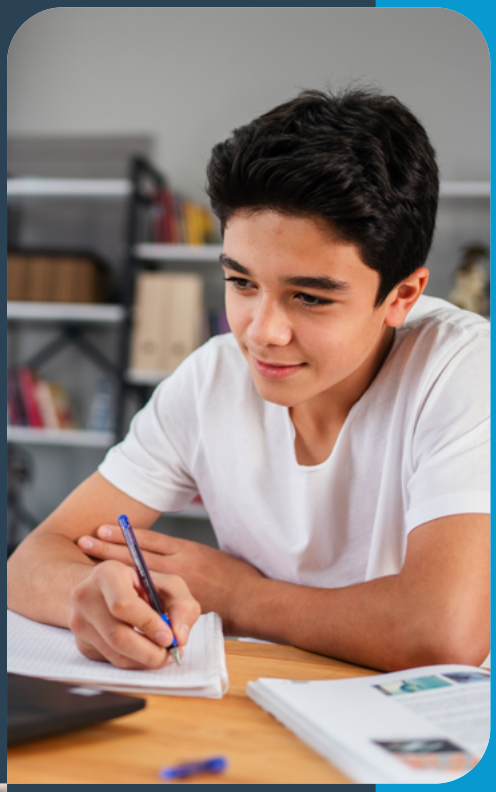
An Introduction to Considerations for Through-Year Assessment Programs:

Purposes, Design, Development, Evaluation

A Paper Prepared for the Smarter Balanced Assessment Consortium



Nathan Dadey & Brian Gong
Center for Assessment

February 3, 2023

smarter BALANCED
Beyond Standard

Acknowledgements

This paper was made possible only through the sponsorship by Smarter Balanced Assessment Consortium. Smarter Balanced provided generous financial support but allowed complete freedom for the authors in terms of content. Smarter Balanced arranged a very efficient review from several experienced state department employees from the Smarter Balanced consortium states who represented diverse backgrounds and perspectives. Their comments helped improve the content, organization, and accessibility of the paper immensely.

The Center for Assessment also has supported the authors' work on through-year assessment and related topics over the past several years. In particular, the Center for Assessment sponsored a convening in 2021 around the topic of through-year assessment (National Center for the Improvement of Educational Assessment, 2021). Many of the ideas build on in this paper were developed for that convening. We also profited quite a bit from the insightful comments from many of those who attending the 2021 convening.

We also gratefully acknowledge the suggestions and contributions of many, especially the insightful reviews of Carla Evans, Charlie DePascale and Catherine Gewertz. Of course, any errors or shortcomings in this paper are the responsibility of the authors.

Suggested citation:

Dadey, N., & Gong, B. (2023). *An introduction to considerations for through-year assessment programs: purposes, design, development, evaluation [Research report]*. Retrieved from Smarter Balanced website: https://portal.smarterbalanced.org/library/en/2023-sb-consideration-of-technical-issues.pdf

Table of Contents

Table of Contents (continued)

Introduction

State summative assessments have been a fixture at least since the 1990's as a requirement for receiving federal Title I funds. Recently there has been much interest in "through-year assessments," which aim to fulfill the summative purposes required by federal law, *and* address an additional purpose or purposes, such as reducing overall testing time or providing more timely instructional information. Designing and implementing a high-quality assessment that serves a single purpose is challenging, and so it is not surprising that designing and implementing through-year assessments that fulfill multiple purposes well has also proven to be very challenging.

From our years of supporting those interested in improving summative assessments by possibly using through-year assessments, several points have been clear: There is no one "through-year assessment"—there are many possible designs. Among the possible through-year designs, each user's context and purposes will guide which design is more suitable. Each design brings its own challenges and trade-offs. There are significant challenges to developing a through-year assessment system. No one has come up with a design that has been widely accepted as meeting educational, political, technical, and feasibility constraints. Efforts to develop state through-year assessment programs may be characterized as exploratory and developmental, rather than as something with well-accepted solutions and well-understood procedures and costs for implementing. Through-year assessment programs require both much greater tailoring to context and much greater investment of resources than typical state summative assessment programs, as well as greater resources. As of the time this document was written in late 2022, only a very few through-year assessment programs had been fully designed, and even fewer had operational data and been subject to thorough evaluation.

This document is written primarily for policy makers and state department of education staff who are considering through-year assessments, as well as consultants and contractors state departments rely on. The document identifies essential things to consider when designing or evaluating a through-year assessment program. This document is not, however, a comprehensive encyclopedia of all possible through-year assessment designs and topics. It is not a handbook for how to design and construct a through-year assessment. It is not a comprehensive review of the relevant literature and practical work relevant to through-year assessments. It offers no "thumbs-up/ thumbs down" verdict regarding particular through-year assessment efforts by specific states or vendors. And it certainly is not a crystal ball regarding what the future might bring to federal Peer Review or state laws. It does provide a firm foundation, however, for considering the "whether," "why," and "what" of through-year assessment design.

The paper is organized into five sections. The first section provides a definition of through-year assessment, the main motivations and purposes for through-year assessments, and the tools for specifying an assessment design, including theories of action, claims, and validity arguments. The second section describes key design aspects that every through-year assessment program must address, and some options for those design aspects that distinguish through-year models. The third section discusses emerging examples of specific through-year assessment designs in terms of their design choices, challenges, and trade-offs. The fourth section provides suggestions for evaluating through-year assessment programs that go beyond current evaluation requirements for state summative assessments, such as federal Peer Review. The fifth and final section provides conclusions and a view to the future.

1 Background, Definitions, and Motivations

Background

The ideas behind through-year assessment programs are not new. These kinds of programs have been subtly promoted by legislators and the U.S. Department of Education since at least the 2010 Race To the Top grant program. The 2015 Every Students Succeeds Act (ESSA) further signaled an openness to through-year assessment by explicitly stating that statewide assessment may "be administered through multiple statewide interim assessments" (ESSA, §1111(b)(2)(B)(viii); see also Dadey & Gong, 2017). This option, however, gained little traction immediately after the passage of ESSA. In contrast, in early 2023 there are now more than ten states and organizations pursuing through-year assessment programs. (See Appendix A for more detail on the history of through-year assessments in the United States.)

This seismic shift in the assessment landscape is likely due to a buildup of dissatisfaction with the perceived usefulness of current statewide assessments – whose results are often viewed as too late and too general – compounded by several surrounding factors, including legislative initiatives, foundation funding, vendor-based disruption, and the pandemic. In terms of the final factor, the pandemic, statewide assessment has already been subjected to a number of disruptions – most notability the complete suspension of testing in the spring of 2020. In light of these disruptions, many states appear willing to consider substantial changes to their statewide assessment programs.

Definition

There are many different through-year assessment designs, but most meet this general definition:

> A through-year assessment program consists of multiple distinct assessments administered across the school year where information from the multiple assessments is (i) combined to yield a summative determination of student performance to support federally required systems of school identification and support, and (ii) used to support at least one additional purpose.

The combination of key features called out in this definition – (1) a set of multiple assessments administered across the year, (2) whose results are combined produce a summative student determination and (3) also support one or more additional purposes – differentiate it from other kinds of assessment, as articulated in Table 1 below. As a helpful heuristic, we suggest that these additional purposes can be characterized as logistical, monitoring or instructional.

These features, and in particular the requirement that the results be combined to produce a summative score, also make it quite challenging to implement in practice. This requirement, as well as our above definition, stems directly from the Race To the Top grant program call, which notes that the results of through-course, which we now refer to as through-year 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). One key disagreement in the field is whether the results must be combined across the multiple assessments. Many emerging programs only use the last assessment administration to produce summative determinations. One might refer to such systems as "summative state assessment with complementary state interims" or even "interim plus summative" programs. However, the field is generally referring to these programs as through-year. We do the same.

Table 1. **Through-year assessment in relation to other types of assessment**

Type of Assessment	Administration	Measurement Target(s) for Each assessment	Intended Use
Balanced Assessment System	• Multiple assessments • Multiple levels of control • Administered at multiple points during the year	• A unique or overlapping part of the content domain • Together the assessments are meant to provide a coherent, comprehensive and continuous picture of student performance	• Serve a diverse set of uses for a diverse set of stakeholders
Through-Year Assessment	• Statewide • Multiple assessments • Administered at multiple points during the year	• Often the depth and breadth of the state content standards, or • A subset of the content domain	• Produce a summative determination to be used in a state's ESSA required system of school identification and support • To support additional purposes, including logistical, monitoring or instructional
State Summative Assessment	• Statewide • Single assessment • Typically administered at the end of the year	• Depth and breadth of the state content standards	• Produce a summative determination to be used in a state's ESSA required system of school identification and support
School or District Interim Assessment	• Typically school or districtwide • Multiple assessments, often parallel in terms of content and measurement targets • Often administered in Fall, Winter and Spring windows	• Often the depth and breadth of the content domain, as defined through a vendor defined blueprint or framework, or • A subset of the content domain	• Serves a variety of instructional and evaluative purposes, often at the classroom, school and district levels
Classroom Summative Assessment	• Typically classroom or schoolwide • Multiple unique assessments • Administered immediately after instruction or at the end of the year	• Often the content just instructed in the preceding lesson(s) or unit(s)	• Produce grades for individual assessments and be aggregated into course grades

Additional Purposes

The above definition explicitly calls out the additional purpose or purposes that a through-year program is meant to address. Without this additional purpose or these purposes, there is no reason to move away from a typical end-of-year summative assessment. The core motivation behind through-year assessment programs is to accomplish "something else" while also creating the summative determinations required for school identification and support. Each of these purposes imply a major change in design of the typical state's accountability assessment. Essentially, purposes that currently "live outside" of the state assessment are being layered onto the state assessment.

Often, these purposes are given in broad terms. For example, motivations form key stakeholders include: "can we make state assessment more useful to teachers?", "we're testing too much - can we use the interim assessments in place of the state summative?," "Can we spread assessment across the year to make more time to measure deeper learning and complex skills?", or "can assessment be personalized around student needs, progress, and competencies so that we can assess at any time through the year?". Other purposes could be added to these examples.

These examples above identify purposes that might be addressed by a through-year assessment program, and these purposes are quite different from each other. These differences imply that there would need to be quite different through-year assessment designs. The central point is: to design a through-year assessment system—or to know whether a through-year assessment system is needed at all—the purposes, as well as the intended uses of the scores and users for the assessments must be specified.

The success of a through-year program hinges on how well the program can support the given purposes simultaneously, requiring careful design that starts at a clear articulation of the purposes of the through-year program. In many cases, there will be multiple, potentially contradictory, purposes given by multiple stakeholders for the adoption of a through-year assessment program. Clarifying, and prioritizing, the purposes of a through-year program is critical to its success - any single program can only do so much. Not every purpose can be supported within a single through-year program, so specificity is paramount to ensure that the program can support the given purposes. Focusing on the most important purpose(s) supports design and implementation. The purposes of the through-year program must be explicated in detail, and then translated into program logic (e.g., through a theory of action). Doing so requires, for example, moving beyond general notions of "informing instruction" to specific actions supported by the assessment information (e.g., what actions?, by who?, when?, based on what information?, with what supports?).

Below we discuss three main purposes that characterize much of the discourse through-year assessments - logistical (e.g., increase efficiency by using interims for annual summative determinations), monitoring (e.g., identify what unit, like classrooms, grades, schools or subjects, is in most need of additional support), or instructional (e.g., inform instructional next steps within a specific instructional sequence, progress monitoring)

– which act as starting points for these kinds of in-depth considerations. These categories are meant to be a helpful heuristic, rather than a comprehensive set of potential purposes. Considering purpose from other perspectives (e.g., Crane, 2010; Perie et al., 2009; Salvia, Ysseldyke, & Bolt, 2011) may also be useful. Regardless of the particular framing around purpose, any through-year design work will need to articulate the intended uses with much greater specificity, and in doing so answer the question "what are people supposed to do with the results?," in addition to supporting a state's accountability system. Finally, these categories presented below are intentionally ordered from least to most challenging, with instructional purposes likely posing the greatest difficulties to design and implementation.

Logistical Purposes

Logistical purposes are typically concerned with making assessments more logistically feasible or with reducing the "footprint" statewide assessment. The original Partnership for Assessment of Readiness for College and Careers (PARCC) assessment used a through-year design to make the inclusion of complex performance tasks more logistically feasible. The main PARCC assessment consisted of multiple choice and short answer questions administered at the end of the school year, but performance tasks requiring student extended constructed responses were administered earlier in the school year. The earlier administration divided the assessment into more palatable multiple sessions, as the original PARCC design consisted of over ten hours of testing. It also made it feasible to have human raters score the performance tasks by the end of the year. The New Hampshire Performance Assessment of Competency Education (PACE) project, was made up of curriculum-embedded performance assessments, also spread the assessments throughout the year for logistical as well as instructional reasons (Marion, & Leather, 2015).

Another commonly mentioned logistical purpose is to reduce testing time. Thus far the total testing time for current through-year assessment programs appears to be greater than for typical state summative assessment, so reductions in testing time likely need to be considered in a more wholistic fashion. Specifically, testing time could be reduced by replacing multiple separate district interim and state summative assessments with a single through-year assessment program. By eliminating two prior sets of assessments, through-year assessment programs may be able to both reduce testing time and also provide assessments that are more coherent with one another.

In other cases, total testing time is not the issue per se, but rather the fit of the assessment program within the school year. In some cases, a through-year assessment program might be less disruptive to school schedules, particularly if the through-year assessment fits into a single class period and is supported by flexibility in administration. Other logistical purposes deal with student experience. In particular, the multiple administrations of a through-year program might help alleviate student anxiety, making the assessments feel more like part of typical instruction and less like a single high-stakes event. Another subset of logistical purposes deals with providing students with multiple opportunities to test (e.g., multiple opportunities to demonstrate proficiency).

Monitoring Purposes

Monitoring largely deals with decisions around the differential allocation of support or other resources. Often, administrators and policy makers are trying to determine where support is most needed and how it should be provided, given limited resources. Is it a particular grade? A set of classrooms? A set of schools? This kind of formative evaluation, ideally, allows administrators to provide support throughout the year to students and teachers to improve student learning. These supports may include teacher coaching, student tutoring, or organization of professional learning communities. Whatever the strategy, this formative evaluation requires asking: What is the need? What is the best possible solution? Is the solution working as intended? Administrative attention to reducing barriers and improving positive levers is a principal way that student learning may be *systemically* improved.

Through-year assessment results may also be used within summative evaluation to identify and improve the effectiveness of curriculum, instruction, and educational programs and policies. For example, using measures of within-year growth with the usual caveats about causal inference, an administrator might ask "did the two math curricula result in equal gains in student learning?". Student achievement may also be used to contextualize the program evaluation (e.g., "Is the curricular program equally effective for students who started with different achievement levels in math?").

Instructional Purposes

Instructional purposes are those that deal with how a through-year assessment program's design and results support the instructional practice of educators. Since through-year assessment programs involve multiple administrations throughout the year, there appears to be a widespread expectation in the field that results will be reported in ways that support instruction. That is, there is typically an implicit assumption that through-year assessment programs can and will support instruction directly. Therefore, any through-year assessment program will need to attend to instructional purposes, even if is to clarify that the primary purposes are not instructional in nature.

Designing a through-year program to support an instructional purpose, or purposes, involves defining the specific actions to be taken by educators based on the results of each through-year assessment administration.

Designing a through-year program to support an instructional purpose, or purposes, involves defining the specific actions to be taken by educators based on the results of each through-year assessment administration. Ideally, this requires moving beyond general notions of "informing instruction" to specific actions supported by the results for specific users. Research from both formative and interim assessment can help shape the

articulation of instructional purposes. Abrams, McMillan and Wetzel (2015), for example, note that in the context of interim assessments, educators made a variety of broad-based instructional strategy adjustments, including modifications to whole class instruction, working with students in small groups, and providing individualized support. Each one of these kinds of adjustments can and should be unpacked when articulating an instructional purpose. For example, what specifically does modification to whole class instruction entail and for how long? Or what misconceptions will be addressed in follow up lessons that commonly occur for students, based on a typical instructional sequence? Or how will formative assessment conversations be informed based on the through-year results? Addressing these kinds of questions is quite challenging, as both interim and state summative assessments have generally only been distally connected to direct instruction. However, bridging these kinds of gaps will be critical if through-year assessment programs are to meaningfully inform instruction. Like all assessments, through-year assessments must be designed for intended uses, and the more specific the intended uses are, the better aligned the score interpretations and subsequent score reports will be.

An Aside on Interim Assessments and Balanced Assessment Systems

One logistical purpose for the development of through-year assessment programs is efficiency: the replacement of local interim assessments with state through-year assessments. One often heard request from the field is "Can't we just use our interim assessment instead of the state summative assessment?" Because many schools and districts use interim assessments, the question has naturally arisen whether the state assessments might be made more like these interim assessments. This again reflects the goal of consolidating purposes within a single assessment and assessment program, instead of having a diverse set of assessments that work together to meet multiple purposes (e.g., keeping interims outside of state assessment or implementing a balanced system of assessments). Likely, much of the drive towards replacing one end-of-year summative with through-year components stems from perceived lack of utility in state assessment as well as perceived disconnects between summative and interim assessment results.

There are at least three key questions when considering through-year from this perspective (see Dadey & Gong, 2017 for more detail): (1) does the interim assessment program *currently* result in the intended outcomes while minimizing unintended outcomes?, (2) can the interim program be *modified* to support annual determinations of student proficiency used in a high-stakes school accountability context, i.e., produce summative scores?, and (3) after being modified, can the program still result in the intended outcomes while minimizing unintended outcomes? If the answer to any of these questions is "no", then a state may be better served by moving towards a balanced system of assessments (e.g., Pellegrino, Chudowsky, & Glaser, 2001) in which separate assessments are used to meet multiple and often diverse purposes, albeit with greater coherence amongst the assessments.

2 Theories of Action and Validity Arguments

Through-year assessment programs will need to meet all the requirements of current state summative assessment programs[1] in addition to requirements imposed by the additional purpose or purposes. Essentially, through-year designs require state assessments to do double duty by meeting multiple purposes. Doing so means that the state assessment program, in the form of a through-year assessment program, will need to provide new or additional information, to a more diverse set of stakeholders, with a greater degree of support for implementation—yet meet the same technical quality requirements. Typically, assessments are designed for a very narrow set of purposes - often just one (i.e., to support school identification and support). Even with one purpose, as is the case with current statewide summative assessment, the challenges are often sizable. Through-year assessment programs, by seeking to integrate multiple purposes, increase development and implementation challenges, potentially to the point of failure. It is therefore critical to pressure test a through-year assessment program critically and frequently. This is not to say that any given through-year assessment program cannot be successful, but rather that success hinges on attending to the complexity inherent in supporting multiple - and potentially competing - purposes.

Doing so requires a clear articulation of the logic and evidence supporting the through-year program. Following Bennett, Kane & Bridgeman (2011), we suggest that these programs do so by drawing on two key frameworks:

1. A **theory of action** that details the *how* of the program, by articulating:

   a. the inputs into the program, like score reports, additional personnel, or funding;

   b. the actions that various people do in the program, like adjust instruction or reallocate resources;

   c. and the outcomes, like reduced achievement gaps or improved student achievement, based on the additional purposes.

2. An **interpretive and validity argument**, that details the *what* of the assessment, that is what claims are going to be made about what students know, as well as the logic and evidence that supports those claims.

[1] Assuming federal law and Peer Review guidance does not change.

The theory of action framework is widely used throughout work on educational interventions and program evaluation, whereas interpretive and validity arguments (e.g., Kane, 2006; we will refer to both interpretive and validity arguments simply as validity arguments) are widely used within educational and psychological assessment. However, only recently have these frameworks been formally drawn together and used to characterize assessment programs. Doing so helps acknowledge that an assessment program is bigger than just the assessment itself - that reaching any particular outcome through an assessment program requires inputs and actions that go beyond the assessments and its results.

These two key frameworks will need to be developed through joint work by stakeholders, states, vendors, and other parties for each state's through-year assessment program. Each state context is unique, so there is no one theory of action or validity argument. Instead, they will need to be developed to address the unique challenges and contexts of a given state. This tailoring to context is needed even when a state is adopting a through-year program that has already been developed. Some of the already developed logic and evidence can support the adopting states work, but much will need to be developed anew.

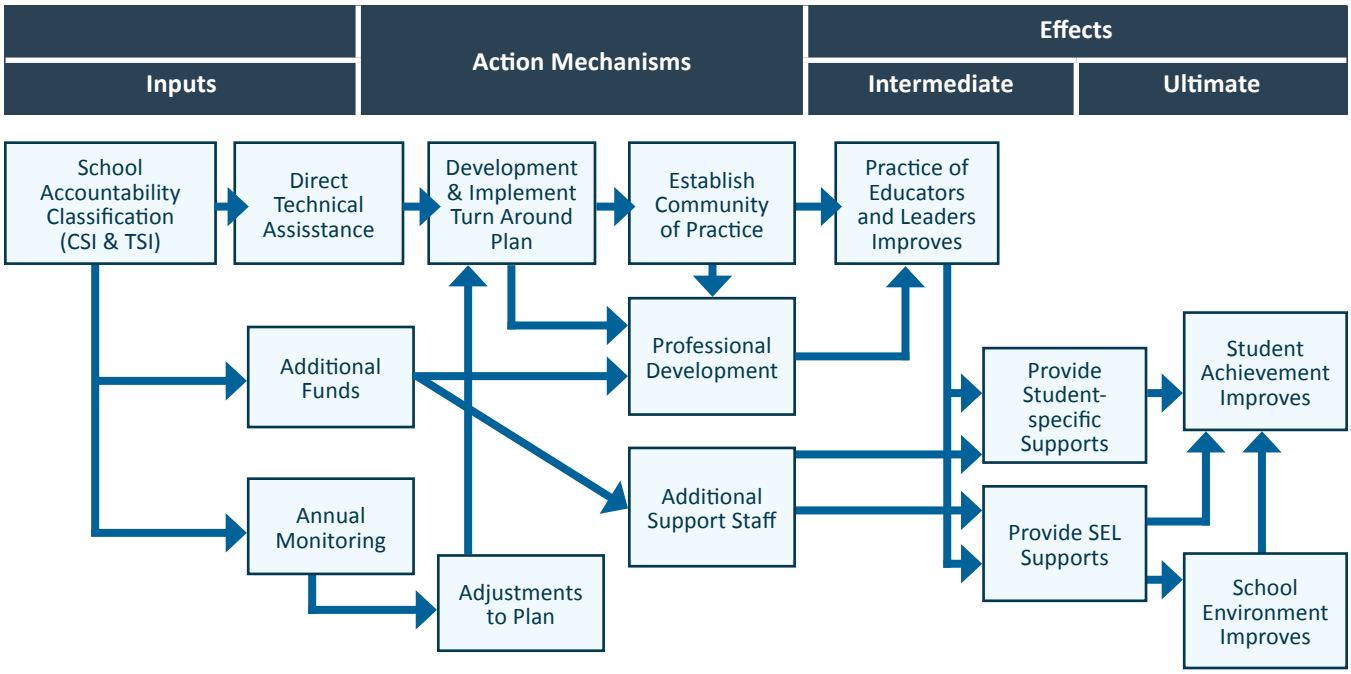
Theory of Action

A theory of action details the inputs, action mechanisms and outcomes of a program. For a through-year program, the theory of action must encompass the program's multiple purposes: the federally required systems of school identification and support, as well as the other purposes. Doing so means that theory of action, or perhaps multiple theories of actions connected together, will need to contend with the potentially divergent inputs, actions and outcomes that stem from the summative and additional purposes. While current state assessments are designed to meet federal requirements for school identification and support, they do not fulfill other desired purposes well (e.g., providing instructionally useful information). The key challenge in designing through-year assessment programs is meeting these federal requirements while also supporting these other desired purposes. Essentially, through-year designs require state assessments to do double duty by meeting multiple purposes. Through-year designs also require an expanded theory of action that explains the key assumptions, mechanisms, and elements that must be in place for a through-year program to serve multiple purposes.

In many cases, the federally required systems of school identification and support and the additional purpose or purposes operate at different time scales, with different inputs, actions, and outcomes. School identification and support, which we often refer to as the summative purpose, operates on a multi-year timeframe. The theory of action for this summative purpose involves first identifying schools for support based on multiple years of data, which triggers supports that play out over a number of years (e.g., the development and implementation of a school turn around plan coupled with additional funds and annual monitoring). Notably, the specifics of this theory of action are left to

states. Federal requirements (as articulated in ESSA), at a high level, only require that states have (1) rigorous content standards, (2) standardized statewide assessments that provide annual determinations of student proficiency and (3) an accountability system that identifies schools in need of support and then provides support. States define the standards, assessments and accountability systems, and everything that goes along with them. A hypothetical theory of action for a state accountability program is given below in Figure 1.

Figure 1. **High Level Theory of Action for a Hypothetical, Example State Accountability Program.**

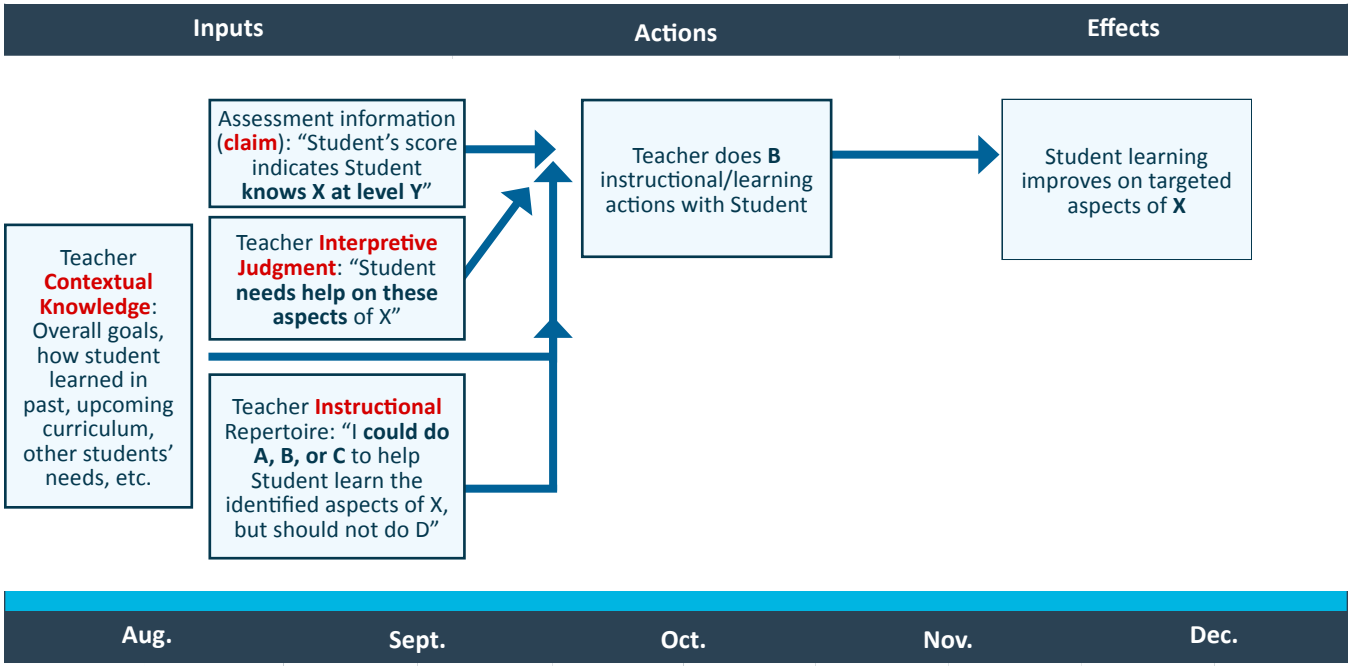


Note: This theory of action is a simplified version of one state's ESSA plan and is provided as an example.

A critical question for through-year assessment programs is how well any additional purpose, and corresponding theory of action, can be layered onto a state's already existing system of school identification and support. Whether, and how, the summative and additional purposes work together, or not, will in large part determine the success of any given through-year program. Logistical purposes require theories of action that are relatively straightforward extensions of the theory of action in place for school identification and support. For example, reducing the footprint of assessment by replacing interim assessments with a through-year assessment program requires a theory of action that details the current uses of interims, whether those uses can be supported by the through-year assessment program and whether the field will move away from current interim assessments. Theories of action for monitoring purposes are more involved, but can still be seen as extensions of the school identification and support theory of action. For example, using an assessment early in the year or growth across the year to allocate resources requires a theory of action that details what these resources are, how they are allocated, and how they are meant to lead to better outcomes.

Instructional purposes, on the other hand, involve theories of action that are substantially different than those currently in place. These differences involve every aspect of the theory of action, including the inputs, actions and outcomes. In terms of inputs, whereas assessment results that inform the school identification and support are intentionally broad measures of state standards that are generally reported after instruction, assessment results that inform instruction must generally be fine grained and connected to ongoing instruction. One key challenge here is that the scope, sequence and pacing of instruction is widely varied across the classrooms, schools and districts within any given state. Therefore, the results that would best inform instruction would likely need to vary from classroom to classroom – a challenge to say the least for statewide assessment. In addition, the actions taken by classroom teachers to improve teaching and learning are also quite broad and varied, so it is an open question as to how a through-year assessment program focused on instructional purposes can detail actions that are sufficiently detailed to help educators know what do to next based on assessment results while also being broad enough to apply to the varied contexts across the state. For example, a theory of action involving assessment results that inform instruction at multiple points during the year would need to detail what time points are most important to provide assessment results at, what the assessment results are, how educators would use those results to adjust their practice and how those actions would improve teaching and learning. This kind of theory of action would also need to address the variation in instruction within the state – the assessment information, timing and actions all may, ideally, need to vary across classrooms, schools and districts. Figure 2 provides a graphical mockup of this example, but still does not provide the specifics needed for an actual theory of action.

Figure 2. **High Level Theory of Action for a Hypothetical, Example Instructional Purpose.**



For any purpose, the key question, is again, how well that purpose's corresponding theory of action works alongside of the state's current theory of action for federally required school identification and support. Given the examples provided in Figures 1 and 2, this question becomes something like: "Can the kinds of specific assessment results that inform instruction be provided by a single assessment program that also provides annual determinations, and if so, do the ways in which teachers act on that information lead to better teaching and learning?" Even this more complex question misses much of the detail that is needed to really consider how well a through-year assessment program might function. Moreover, there is an open question as to whether instructional purposes can coexist with accountability purposes, or whether the pressures of accountability will result in the instructional purposes being minimized through behaviors like teaching to the test (e.g., Campbell, 1976).

A well specified theory of action also considers these kinds of potential threats and approaches to addressing them. Developing a full theory of action for a through-year assessment program is well beyond the scope of this paper. However, the development, testing and revision of a theory of action is the most important part of the development and implementation of a through-year program, as it guides all of the decisions made within the program. Both choices involved in assessment design, like how the content is allocated to the multiple assessments, as well as those surrounding the assessment, like whether and how professional development are provided, are all guided by the theory of action. The first step in developing a theory of action is to be very clear and specific about the intended purposes of the through-year assessment program. It must be articulated *what* information will be provided, *how* it is intended to be used instructionally by *whom*, and *why* that assessment information used that way will likely result in improved student learning. In addition, careful detail will need to be given to whether and how the summative and additional purpose or purposes can work together – in some cases they might not. In these cases, a through-year assessment program would need to either be reconceptualized or abandoned. Finally, developing a theory of action is an involved endeavor, ideally involving much iteration. Tools like Logic Models (e.g., Frechtling, 2007, W.K. Kellog Foundation, 1998) or Driver Diagrams (e.g., Bennett & Provost, 2015) can be quite helpful in developing theories of action.

The first step in developing a theory of action is to be very clear and specific about the intended purposes of the through-year assessment program.

Interpretive and Validity Arguments

Through-year assessment programs aim to meet multiple purposes. To support the multiple purposes of a program, multiple scores will be needed that provide information about different aspects of student performance. Ideally, the theory of action helps define what kinds of scores are needed, as well as how they are used. All through-year assessment programs are required to produce a summative score that is used in state systems of school identification and support. Another type of score or scores are needed to support the additional purposes. These additional scores will likely vary from one through-year assessment program to another. These scores, for example, may be overall summaries of student performance at multiple times during the year, scores that quantify growth across the year, or more fine-grained scores on specific knowledge and skills at multiple times during the year.

Key is defining what claims are to be made about students. For example, consider two hypothetical claims:

- A summative claim: "The student was proficient on the knowledge, skills and abilities represented by the state's college-ready, grade-level content standards at the end of the year."
- An instructional claim: "The student needs additional support to master concepts they have recently been instructed on."

Each of these two statements, or claims, are meant to encompass the core idea about what is going to be said about students know and can do. There is no one agreed upon way to formulate a claim, but we find it helpful to state them as single sentences. In practice, a great deal of additional detail is needed to support the development and implementation of a through-year assessment program.

Even though these two statements are not as nearly detailed enough to support the development and implementation of a program, these two claims clearly refer to two different bodies of content: the first, the summative claim, is about students' proficiency on the grade-level content standards whereas the instructional claim is about students' mastery of what they have recently been instructed on. An important complication is that "what they have recently been instructed on" may not be grade-level content, nor is it likely to be the same content from classroom to classroom. For example, at the same point during the school year, in one classroom students may have just wrapped up instruction focused on early multiplication and division whereas another classroom may be in the middle of instruction on understanding fractions (e.g., Cole & Swanson, 2022). The claims being made will either need to be broad enough to accommodate this kind of variation, or the design of the through-year assessment program will need to be flexible enough so that the right content can be administered at the right time to support the right claims.

In sum, the hypothetical examples make clear that the summative claim is about students' proficiency on the grade-level content standards, whereas other claims, like a claim rooted in instruction, may involve content that is off grade-level and varied from classroom to classroom. The design, including the theory of action, of a through-year program will need to navigate this issue to meet both summative and additional purposes.

Summative Claims

Producing a summative score to support state systems of identification and support requires defining what, specifically, is going to be inferred about what students know and can do. Whereas doing so is relatively straightforward for typical state summative assessment programs, doing so for a through-year assessment program requires considering not only *how* the state content standards will be assessed, but *when*. Since through-year assessment programs are administered through a set of assessments spread out across the year, the timing of the assessments becomes an issue. Consider the two claims below:

- A typical summative claim: "The student **was proficient** on the knowledge, skills and abilities represented by the state's college-ready, grade-level content **standards at the end of the year**."
- An alternative summative claim involving both end-of-year and within-year results: "The student was **typically proficient** on the knowledge, skills and abilities represented by the state's college-ready, grade-level content standards **at multiple points during the year**."

The second claim changes how proficiency is assessed and in doing so also changes what counts for accountability – instead of determinations being based on results from a single end-of-year assessment, now results from multiple assessments administered during the academic year now contribute to annual determinations. This kind of change involves addressing technical issues (e.g., how to create a "single summative score"), but these issues pale in comparison to the policy issues involved. Technical solutions can be developed or adapted from previous solutions. Although there is much unknown about supporting summative claims based on within-year information, these unknowns can be addressed through careful research. There are only a few overall types of approaches that support these claims, and hopefully a clear picture of the implications of these approaches will emerge in the upcoming years. In contrast, policy solutions may be much harder to come by. Successfully navigating the policy issues involved in making summative determinations based on both within-year and end-of-year results, and the highly variable patterns of teaching and learning across the state that through-year assessments are embedded in, involves considering not only what to assess and when, but the values of the variety of stakeholders invested in statewide summative assessment.

Technical solutions can be developed or adapted from previous solutions. Although there is much unknown about supporting summative claims based on within-year information, these unknowns can be addressed through careful research.

Much of the complication arises from challenges involving (1) how ongoing teaching and learning, and presumably improved performance across the year, are addressed in relation to the summative claim and (2) how the claim, and its implementation in a through-year assessment program, interacts with the widely varied scope, pace, and sequence of instruction across the state.

In terms of the first challenge, the issue is that students may do better if they been assessed at the end of the year. Or more generally, the challenge is that students may perform differently – better or worse – if they had been assessed at other points in time during the year. Supporting a summative claim, and creating corresponding single summative scores, based on results from multiple points in time may produce scores that are generally lower than those produced using just end-of year assessment results. Complicating this general trend is that while many students may receive lower scores relative to those based on the end of the year results, some may not. So, the question is not only whether scores are lower, but lower for whom.

However, these considerations assume that student performance at the end of the year is the *right* frame of reference. A state might decide that a claim based on results from within the year is valued, regardless of whether students might perform better based on results from the end of the year. Restated, changing the summative claim is primarily a matter of policy. A matter tied up in what the state wants to say about what students know and can do, as well as how the state's varied stakeholders understand and react to that claim (e.g., whether a given summative claim is fair).

In terms of the second challenge, changing the claim to one involving within-year results introduces complexity related to variation in the scope, sequence and pacing of instruction across the state. The typical end-of-year summative claim does not have to address *when* students are instructed on specific knowledge, skills, and abilities, since students have presumably received the sum total of instruction. A claim based on results from within the year does. Supporting this claim involves considering the messy interactions between what is assessed, when it is assessed and the patterns of teaching and learning in the state. For example, some students might be assessed on content they have not yet been instructed on.

Whether this is the case is a function of both instruction and assessment design. A through-year program might well have a design that tests students on untaught content, or, with careful planning minimize such occurrences to a great degree. Doing so may involve changing both the summative claim as well as the supporting assessment design. Consider the following claims:

- "The student was **proficient** on the knowledge, skills and abilities represented by the state's college-ready, grade-level content standards **at least one point during the year**."
- "The student was **proficient on small groupings** of the knowledge, skills and abilities represented by the state's college-ready, grade-level content standards shortly **after instruction related to those groupings**."

Each of these claims has implications for the design of a through-year assessment program, including the design of the assessments, the construction of the single summative score, and assessment administration, as well as the relationships the program has with the scope, pacing and sequences of instruction across the state.

Each of these claims also reflects a specific set of values and priorities, which inform the design of a through-year assessment program. Which of these claims can and should be used as the basis of a through-year program? The current answer now is “all of them”. There are multiple through-year programs emerging, each of which has adopted a claim rooted in one of the above examples. Notably, many programs are maintaining an end-of-year summative claim by only using results from the last assessment, while a number of other programs are designed to support claims that incorporate within-year results.

Whatever the claim, it will need to be supported by logic and evidence, typically in the form of an interpretive and validity argument (Kane, 2006). This approach is widely accepted and can be applied to through-year programs. Key in informing this argument is clarity around what is being inferred about what students know and can do. This clarity is also critical in informing decisions about the design of the program itself, which we turn to next.

3 Key Technical Design Features

Designing any assessment program involves a whole host of decisions, each of which is accompanied by its own body of literature and accompanying practices. In this section our goal is not to detail every possible aspect of through-year assessment design, but rather to address important features that pose unique challenges in light of through-year assessment programs. Each of the listed features is presented individually below, but they are all interconnected. For example, it is difficult to consider the design of the assessed content without considering the structure of the administration that supports it. These key features are: (1) Content Structure. How the content domain like English Language Arts or Mathematics is organized – allocated, divided, or otherwise articulated - across the multiple assessments, (2) Administration. How the multiple assessments are administered, in terms of windows, order and flexibility, and (3) Aggregation. How annual determinations, or what we will refer to as a single summative score, are created. Further detail on these features are provided in Table 2 below.

Table 2. **Summary of Key Design Features.**

Content Structure	Administration	Aggregation
• How the content domain is **organized or structured** across the assessments, which helps define • The **number** and **timing** of the assessments as well as, • The **grain-size** at which the content is allocated to each assessment.	• Whether the assessments are administered in **windows** or **on demand**, as well as • Whether the **order** of the assessments is fixed or flexible, • Which assessments are **required**, and finally, • **Who decides** which assessments are administered and when.	• Whether the single summative score is based on **both within-year and end-of-year results**, or only **end-of-year results**, which is informed by, • **State values** and the **summative claim**, and • Supported by a **measurement model** and **score creation** method.

These features also help differentiate between emerging through-year models. Although there are over a dozen states currently developing or considering through-year assessment programs, these efforts can be sorted into five major models using the above features, as we detail in the *Examples of Possible Through-Year Assessment Designs* section. Within the description of each of these features, we invent language to better describe the various options. We hope that this language helps clarify, rather than hinder, the emerging work, as terminology has been inconsistent within the field. Finally, these features are important, but not the only features worth investigating in relation to through-year assessment. In particular, …

Content Structure

One of the most important features of a through-year assessment program is how the content domain (e.g., the domain of English language arts/literacy (ELA/L), mathematics, or science) is organized across the multiple assessments (i.e., how the content domain is structured). The way in which the content domain is structured is a central design decision that affects every other design decision involved within a through-year assessment program (the number of assessments, their administration, etc.). What is assessed, and when it is assessed, affects every other aspect of a through-year assessment program. The utility of the assessments, particularly for instructional purposes, is tightly bound to the way in which the content domain is structured across assessments. For instructional purposes, in particular, assessing the right content, at the right time, is critical when providing actionable feedback to the field[2].

Prior work has typically described the domain structure in terms of the blueprint (and in particular whether it is the same or different across assessments, see Dadey & Gong, 2017; Gong 2010; Gianopolus, 2019 & Wise, 2011) which falls short to capture the full range of possible models, as blueprints are typically aimed at just a handful of relevant features (e.g., standards, cognitive complexity) whereas the domain can be structured across assessments using these features, or less common features that may be theoretically or pragmatically important (e.g., learning progressions, curricular units). How the content domain is defined, divided, and articulated across the assessment "modules" is based on:

- the aspect or "structure" of the content domain that is used to allocate content to each module (e.g., are the divisions based on the standards, progression or curriculum),
- the grain-size at which content is allocated to each assessment module and
- the flexibility, or lack thereof, in the ordering and timing of the assessment modules.

Decisions about the domain in turn drive the number of assessment modules and how they are administered. Below we first name and then describe possible approaches to structuring the content domain, moving from relatively simple to quite complex. These approaches are meant to be illustrative, not exhaustive; there are likely far more approaches than we present below. Providing an adequate description of these approaches involves blurry terminology, so we introduce and define the term "module" to provide clarity and distinction amongst the designs. Specifically, we use the term module to indicate what would be typically thought of as individual "assessments". However, under some designs a module only covers a small part of the domain, meaning that the assessment program is made up of numerous modules. Finally, note that these approaches to content alone do not fully define any particular model of a through-year assessment program, instead doing so involves choices around other design features, including score aggregation with the associated implications on test security and administration.

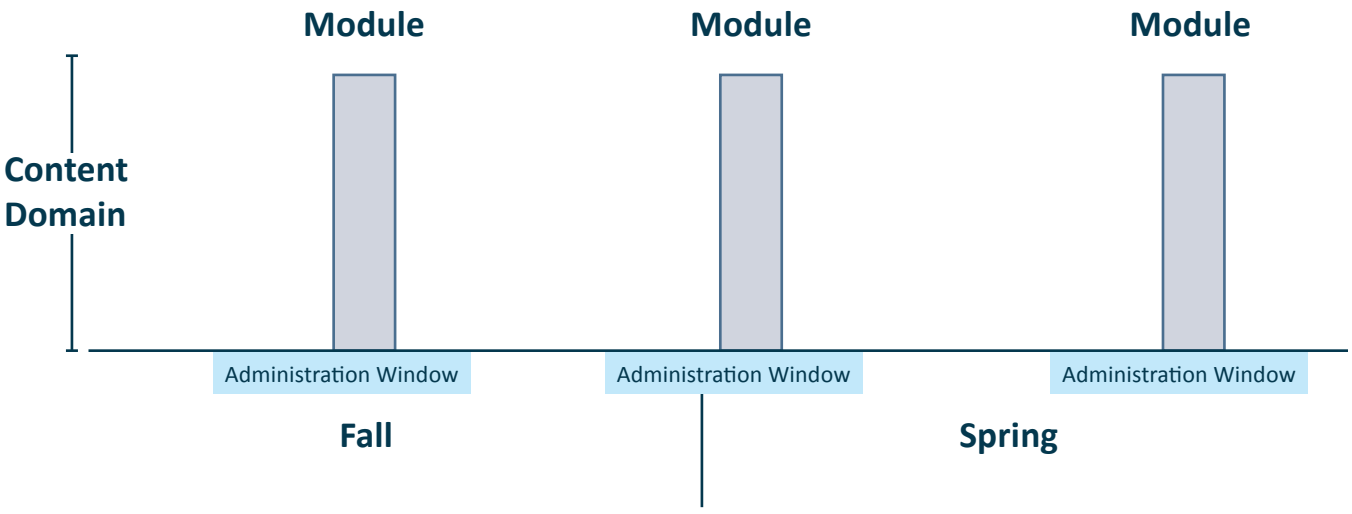
[2] Even then, providing actionable feedback to the field remains an extremely challenging task.

"Full Domain" Approach

One possible approach to structuring the content domain, currently being considered or implemented by multiple states, is to have each module cover the entire content domain (i.e., match an overall blueprint with the same standards coverage at the same level of rigor or complexity). Essentially, each assessment is a parallel version of one another. This model has also been referred to as "full scope" or "mini-summative" design, although the term mini-summative can be misleading, as these assessments are often not shorter than a typical end of the year summative, although testing time can be reduced through the use of an adaptive design (either item level adaptive or multi-stage adaptive). Under this approach, each assessment is either exactly the same or parallel forms of one another, meaning there is no need for articulation across assessments and ordering is a non-issue. This approach is generally implemented through an adaptive design within each assessment administration, or by treating each administration as a single stage of a multistage-adaptive testing program.

Assuming three fixed assessment windows, such an approach might look like Figure 3 below. This figure presents a heuristic for understanding the content domain by test administration. Along the x axis is time, and along the y is the content domain. Each bar represents a distinct administration of the assessment (i.e., three administrations). In this figure, the content domain is the same, as indicated by the same height and shading of the bars. For clarity, instead of referring to each as an assessment, we refer to each unique assessment as a module. Under this approach, we have the same module administered three times within three reporting windows (indicated in blue below).

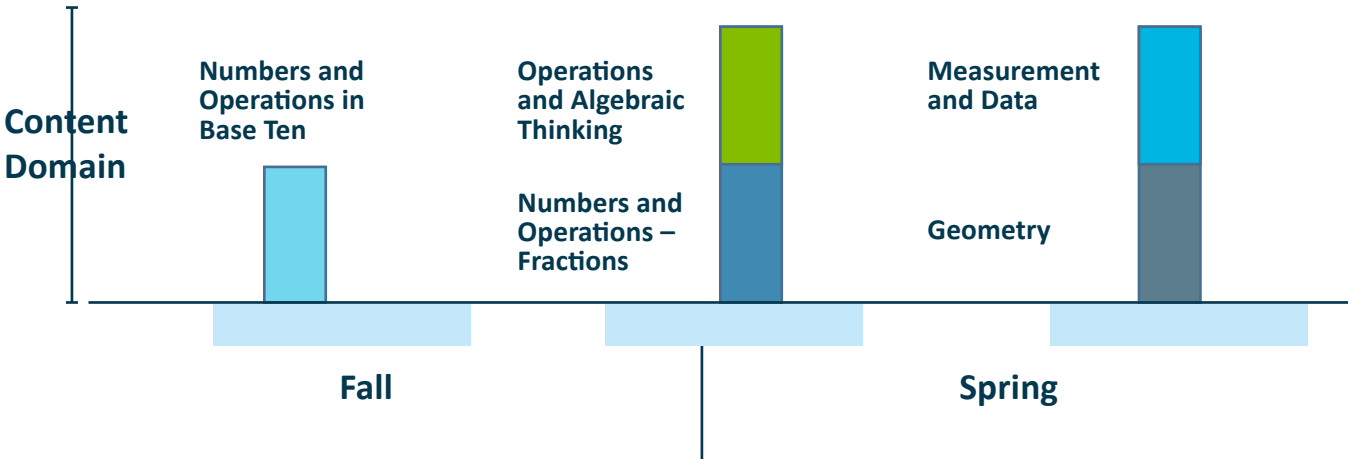
Figure 3. **"Full Domain" Content Approach.**



"Modular Standards Sub-Domain" Approaches

An alternative design to the full domain approach is to divide up the content domain. There are a large number of ways in which the domain might be divided up, be it based on learning progressions[3], curricular units, level of complexity or rigor, or standards. Much of the work around dividing up the content to date has focused on the standards, given the importance standards hold in statewide summative assessment. One approach to do so is by drawing on the hierarchical organization of a given set of standards. For example, in mathematics the content domain could be divided up using the sub-domain groupings that exist within most standards. Figure 4 provides an example of the sub-domain based approach in mathematics, using the sub-domains from the Common Core State Standards. In this figure, there are three modules, the first of which has a single domain assessed, and the latter two have two domains each.

Figure 4. **"Modular Sub-Domain" Content Approach.**



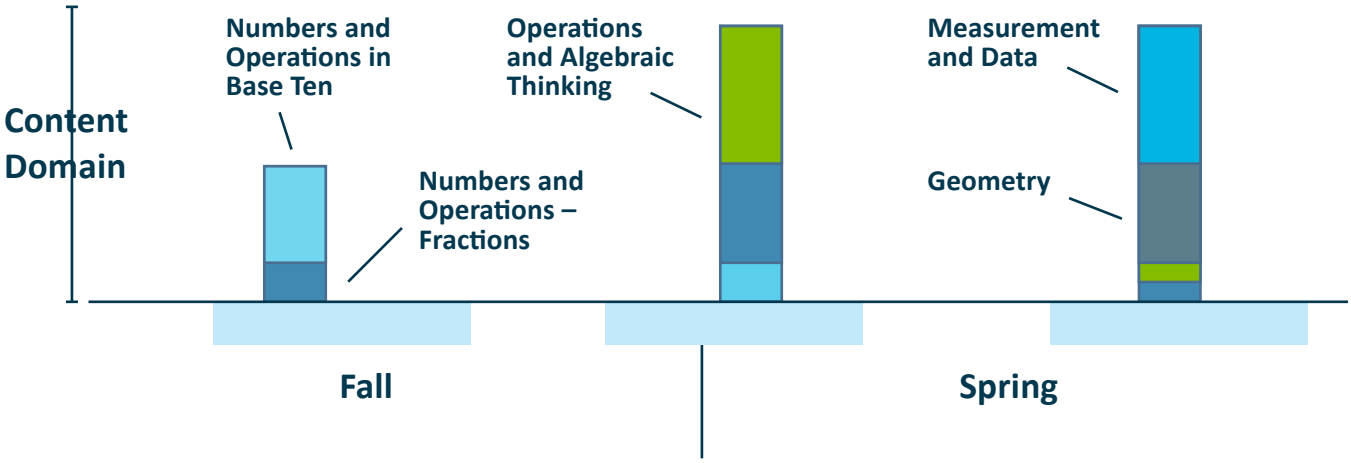
Dividing up the domain in any fashion entails a number of key decisions, as previously outlined above: the aspect or "structure" used to allocate content, the grain-size of the content and the flexibility in ordering and timing of the modules. A key question when the content domain is not fully represented on each administration is whether the division or divisions of content interact with instruction in ways that advantage or disadvantage particular sequences of instruction (e.g., that some students may not have had the opportunity to learn the assessed knowledge, skills and abilities by the time they are assessed). The full domain approach partially sidesteps this issue by assessing the entire content domain on each occasion. Addressing this issue is bound up in what the intended summative claim is, as we address in the next section.

[3] Note that learning progressions do not currently exist for all grades and subjects, nor are learning progressions stable across contexts. Learning progressions are, in some cases, dependent on curriculum and instruction.

The above example provides a single sequence of content across three administration windows. To support this, there needs to be a statewide agreement about how to partition (sequence and divide) the content to be assessed and this has been challenging for many states. A recent study of several popular commercial ELA/L and mathematics curricula found the curricula themselves did not share a common scope and sequence (Cole & Swanson, 2022); each state should conduct or otherwise obtain an analysis of the content organization of the curricula used by districts in the state to determine the feasibility of having agreement about how content ought to be partitioned for a through-year assessment, and the implications if the partitioning does not match some students' opportunity to learn.

There are alternatives that partially address concerns about the relationship between the interaction of the allocation of content and instructional variation. One approach is to partially overlap the content, such that portions of each domain are assessed within each window. For example, the sub-domain approach shown in Figure 4 might be reworked into an "overlapping sub-domain approach" shown in Figure 5.

Figure 5. **"Overlapping Sub-Domain" Content Approach.**



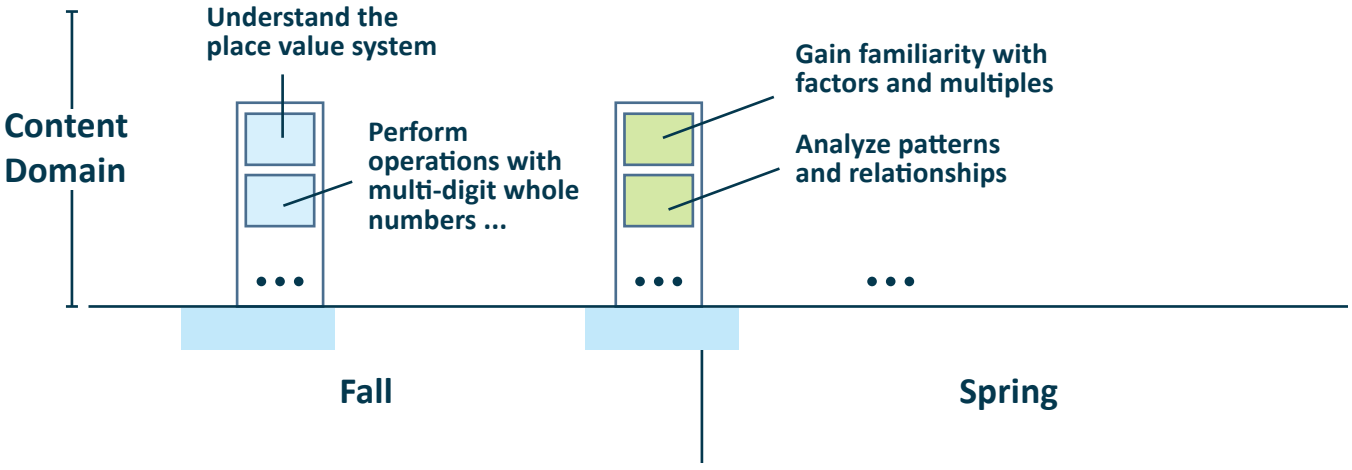
This "overlapping sub-domain approach" could be further extended, becoming a cumulative design in which domains are introduced and retained overtime, eventually building to a final module that covers all of the sub-domains.

Another alternative is to introduce *flexibility* in the order of the administration of the sub-domains. Doing so involves making decisions around the degree of flexibility (e.g., will, say, two options for the ordering of the domains be offered vs. numerous combinations) as well as the level at which decisions are made around flexibility (e.g., the state deciding on the sequences vs. local decision makers, like district leaders, school principals or teachers). A conservative approach to flexibility might involve the state providing, say, two options for administration that vary the content emphases in each window. A less conservative approach, for example, would be to let district leaders submit a plan to the state for the administration of groupings of the sub-domains within three windows. Taken to the extreme, flexibility could allow for individual teachers to choose when to administer each set of individual sub-domain content as its own module at any point during the school year.

"Modular Standards" Approaches

The prior content approach draws on the hierarchical structure of the standards to organize the assessed content. An alternative would be to use the standards themselves directly, by creating modules organized around small groups of standards or individual standards. This approach might be constrained so that it looks fairly similar to the modular standards sub-domain approach. In the example shown in Figure 6 below, standards based modules are grouped together in two distinct administrations. These groupings could be created at the state level, resulting in a single administration order and sequence. The state could, as with the standards sub-domain approach, also create multiple groupings to fit various instructional sequences. Increasing flexibility under this approach could mean that the groupings, and even windows, are developed at the school or district level.

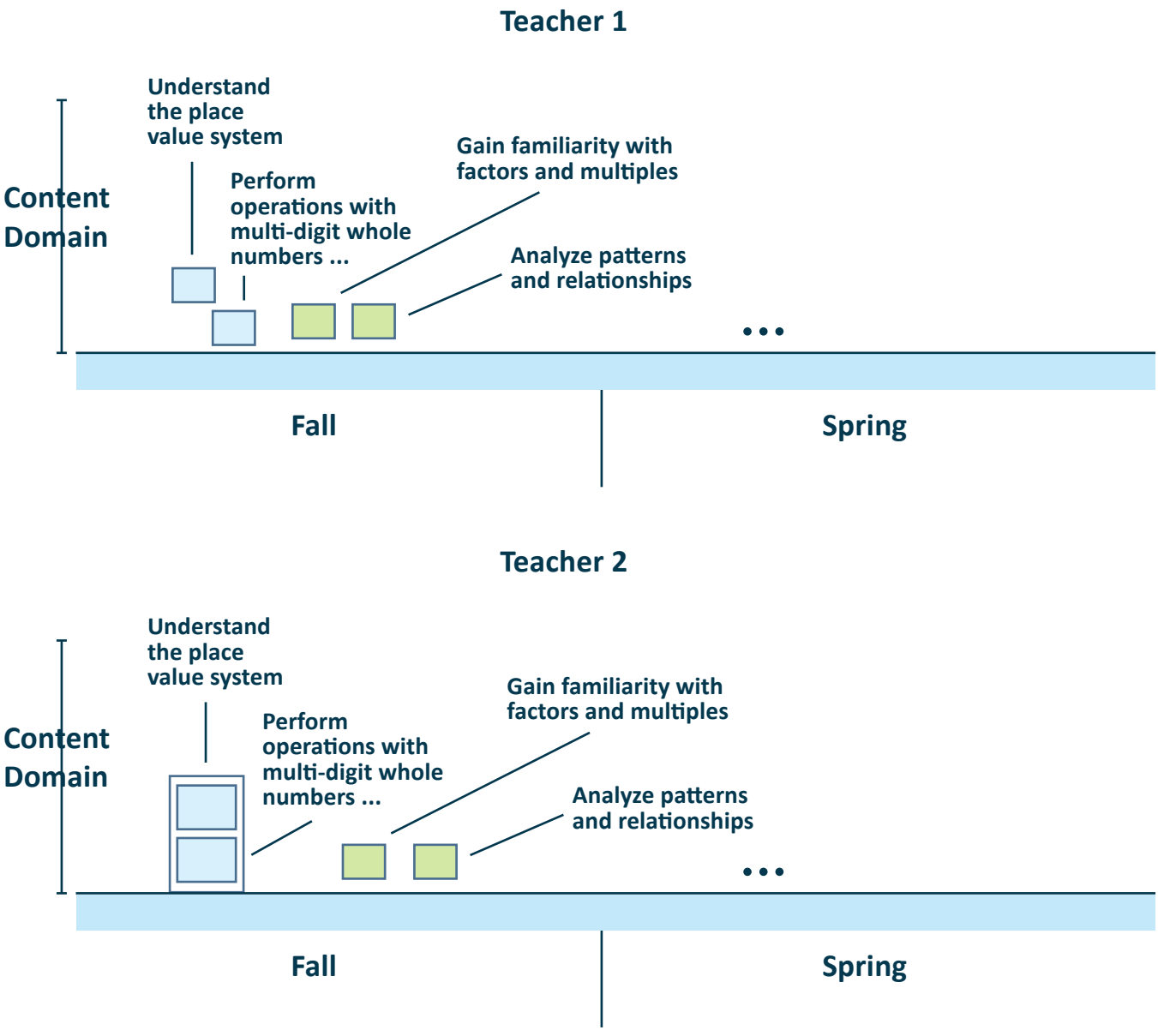
Figure 6. **"Modular Standards" Content Approach with Grouped Administrations.**



Notes: The box around each individual standards module represents a grouping that corresponds to an administration. Only a limited number of standards from two mathematics sub-domains are shown.

Even more flexibility results in, for example, the approach shown in Figure 7 in which teachers can administer modules for individual standards at any point during the year based on their own judgment, subject to the constraint that they administer the full set of modules needed to capture the depth and breadth of the standards.

Figure 7. **"Modular Standards" Content Approach with Flexible Administrations.**



Alternative Approaches

The prior approaches draw on the standards as the primary way in which content is designed and then allocated to modules. However, learning does not occur in discrete chunks that align to standards, rather the knowledge, skills and abilities that underlie any given learning sequence can and do span multiple standards across multiple domains at any given point. For example, students may build towards standards mastery by revisiting components of individual standards several times over the course of a year or grade-band. In addition, standards themselves represent the targets of instruction at the *end of the year* - they intentionally provide no direction into what occurs during the year. Given these considerations, it is easy to imagine any number of additional approaches to developing and allocating content. These approaches could draw their structure from:

- Curricular scopes and sequences, perhaps by either designing modules around commonalities among popular curriculum in the state or allowing for the flexibility to match modules to instruction. Doing so could mean, for example, modules based on curriculum units (and thus sets of modules, each of which is aligned to a specific curriculum).
- Research based learning progressions, in the cases where research based learning progressions are available (e.g., early mathematics) and relatively stable across various curriculum (or having multiple progressions with corresponding sets of modules).
- Models of complexity or sophistication in which knowledge, skills and abilities are articulated in increasingly complex ways up to the standards (e.g., having levels of sophistication for each standard).

These kinds of approaches are not totally incompatible with the prior standards based approaches. For example, the individual standards modules could be provided in sequences that match the sequence of a variety of curriculum. In addition, other considerations (e.g., allowing for retesting of modules; including off grade content) adds additional complexity on top of these approaches.

Learning does not occur in discrete chunks that align to standards, rather the knowledge, skills and abilities that underlie any given learning sequence can and do span multiple standards across multiple domains at any given point.

Connecting Content Structure to Purpose

A hallmark of through-year assessment programs is that they are meant to support both the creation of summative determinations and an additional purpose or purposes. These additional purposes are key in driving the design of the content structure, as without these additional purposes, the best design would likely be a single end-of-year summative assessment. And as we note before, instructional purposes are almost invariably involved within through-year design. Given this, we consider the implications of content design in terms of supporting summative claims and instructional claims.

Content Considerations for Summative Claims. Supporting a summative claim about student proficiency on the state content standards involves considering not only *how well* the assessment program represents the state content standards (e.g., based on evidence of alignment, as well as the overall set of validity evidence), but also *when* the standards are assessed (e.g., considering fairness broadly and, when flexibility is involved, comparability across the unique patterns of modules administered). Criteria for considering these issues is present in the current Peer Review guidance (USDOE, 2018), although these criteria do not explicitly attend to through-year assessment programs.

For programs that take the full domain content approach, implementing an alignment may be a fairly straightforward application of typical alignment methods that involve multiple forms. That is, if each module is counted within the summative score, then each module can likely be considered a parallel form, and subjected to typical alignment methods (and in doing so provide evidence for Critical Element 3.1 from Federal Peer Review). However, considerations of fairness are less straightforward and in ways that are not well reflected within current Peer Review guidance. In particular, assessing the full domain at multiple points during the year, when instructional scope and sequence varies and thus opportunity to learn varies, could result in some instructional sequences being more closely aligned to the assessed content than others, particularly if some parts of the content domain are more heavily emphasized than others (e.g., some content standards sub-domains have more weight than others). There is no technical solution for this issue, although empirical analysis examining the scope and sequence can help. Rather, it is a question of what stakeholders can agree on as appropriate to evaluate students across the school year (i.e., what they value).

For programs that do not take the full domain content approach, considerations of alignment, as well as fairness and comparability, become more complex. In terms of alignment, the full set of modules, together, would need to be evaluated using an alignment methodology. This methodology would need to both consider the total sum of the content, as well as how it is emphasized within the single summative score produced to support annual determinations. The issue with fairness is particularly pointed when the content is divided across multiple modules and *no* flexibility is provided. That is, are some instructional sequences being more closely aligned to the assessed content than others? Will the stakeholders accept such a design? Addressing this question will likely be an area of active research in the future. Providing flexibility in administration may help alleviate this concern, but raises others, including: how do we know any given pattern of modules provides the best representation of what students know and can do relative to the summative claim being made?

Content Considerations for Instructional Claims. To support an instructional claim, a through-year assessment program must produce information educators and leaders find useful to inform instruction day-to-day, week-to-week and month-to-month. This kind of instructionally relevant information is typically fine-grained, connected directly to curriculum and instruction, and content-referenced. Typical state summative assessments, however, are not well suited to providing this kind of information due to restrictions imposed by their use within state accountability systems, as well as limitations due to cost and feasibility. Summative state assessments then, even if administered multiple times during the year in the form of through-year assessment, may not be able to provide the kinds of detailed information produced through classroom-based assessments and formative assessment practices. Whether this is the case depends on how well the through-year assessment program is able to successfully navigate these kinds of restrictions to produce instructionally useful information.

In doing so, those designing a through-year assessment program will need to address tensions around what is assessed, when results are provided, and how the assessment results relate to curriculum and instruction. In terms of what is assessed, instructional utility stems, in part, from the alignment of assessments to the learning targets of students. For state summative assessment, these learning targets are the grade-level state standards. For assessment that provides assessment information that is directly useful to instruction, these learning targets may be directly related to the grade-level state standards but they may not be. The learning targets may be smaller in grain-size, capturing part of a standard, or only partially align to a given standard. Alternatively, the learning targets may be aligned to off grade-level standards. Through-year assessment programs may address these kinds of learning targets, but not exclusively. That is, items that do not align to the grade-level standards could be included in the through-year assessment program, but likely could not be used directly in the production of a single summative score.

Instructional utility also stems from the timeliness of assessment results – when the results are provided. The timeliness of assessment information can range from the minute-by-minute cadence of formative classroom assessment practices, through multiple-times-during-the year interim assessments, to once-a-year state summative assessments, and on to once every multiple years (e.g., the National Assessment of Educational Progress). Critically, what is timely for one student or classroom may not be timely for another, due to differences in the scopes and sequences of instruction. Some content structures, like the modular standards design with flexibility, may be better able to provide results in a more timely fashion. However, the instructional usefulness does not exist in isolation, instead it must be considered in relation to a detailed theory of action. For example, pre-assessment information intended to inform instruction on new content must be provided before instruction on that content begins. Remedial assessment information intended to help identify what a student has been instructed on but has not learned well should be provided as soon as feasible to make a difference and within the feasible instructional cycle.

What is timely for one student or classroom may not be timely for another, due to differences in the scopes and sequences of instruction.

Another aspect of instructional usefulness deals with the through-year assessment programs' relationship with curriculum and instruction. State summative assessments are typically built to be explicitly agnostic to any particular curriculum, but as sensitive as possible to the state content standards. This agnostic approach is rooted in federal Peer Review, which requires that all assessment "forms adequately represent the State's academic content standards and yield consistent score interpretations such that the forms are comparable within and across school years." (Critical Element 4.5; USDOE, 2018, p. 57). Generally, this requirement is interpreted as each form of an assessment within a given grade and subject be as parallel in content and difficulty as possible. This requirement stands in contrast to, for example, creating specific assessments for each curriculum.

Many educators, however, would like assessment results that are more sensitive to their particular curricula and instruction. For example, in checking for student understanding of a recently taught concept or skill, an educator might want an assessment that was highly coherent with the curriculum, in terms of using the same terminology, rooted in the same explanations that students had practice and scaffolding on, and reflecting how particular concepts and skills were developed together. In practice, some through-year assessment designs may be able to strike a balance between the requirements of Peer Review and the desire for connections to curricula. For example, the content of a through-year assessment program might be directly connected to the variation in curriculum and instruction through a flexible content design like the *Modular Standards Content Approach with Flexible Administrations*. Under this approach, educators might be able to select when to assess each standard in line with their instructional sequence, so long as they assess all required standards (see examples of this approach outlined in Dadey, 2018 & 2019; Clark & Karvonen, 2021; Georgia Department of Education, 2018). Alternatively, content might not be directly connected to instructional sequences as in a content design like the *Full Domain* content approach. In this case, more indirect connections to instruction and curriculum could be made through score reports and supports that are curriculum specific (e.g., Dadey & Badrinarayan, 2022). Whether any of these approaches to connecting to curriculum are sufficient to inform instruction is an open question, and one that should be explored empirically.

Ultimately, for a state to design a through-year assessment that provides useful information to inform instruction within the same year and same students as when the assessment was administered, the state should identify what information would support better learning and teaching, and then how that information might be provided. The state will need to determine whether the ideal content organization (e.g., Modular Standards design) for instructional purposes can be made consistent with the content organization for the summative assessment design, and vice versa, given the state's values and priorities. For a state procuring interim assessments as the within-year modules of a through-year assessment, the state will need detailed information about the items, test blueprints, item-to-standards alignment, and score reports to understand the potential for instructional utility and ultimately, construct a interpretive and validity argument.

Administration

Design Decisions

Because through-year assessment designs involve multiple assessment modules that, collectively, serve multiple purposes, the administration of these programs is far more complex than a single end-of-year summative assessment. Decisions about the way in which the models are administered should be informed by the theory of action, and made more concrete through decisions about the content structure. Important decisions on the administration design include:

- Whether the assessment modules are administered within fixed windows or are administered on demand.
- Whether the order of the assessments is fixed or flexible.
- Which assessments are required.
- Who decides which assessments are administered and when.

These design decisions help shape the administration experience of students, teachers and leaders. These decisions can be thought of in terms of a continuum of standardization to flexibility. Towards the standardization end of this continuum, assessment modules are administered within narrow fixed windows in a predefined order determined by the state. Towards the flexibility end of the continuum, assessment modules are administered on demand throughout the year in any order, as determined by teachers. Within these more extreme versions of standardization and flexibility are a variety of options. For example, a state may opt to let local districts define assessment windows and corresponding order of the modules. No state to date has taken this kind of approach, but it is well within the bounds of the current extremes within the field.

These decisions interact with the content structure of the through-year assessment program. For example, ordering does not matter much, if at all, for the Full Domain content structure. The overall experience remains much the same, as the assessment modules parallel one another in terms of content. On the other hand, having flexible ordering paired with a Modular Standards content design leads to a highly variable experience for students and teachers, especially if that flexibility in ordering is coupled with on demand administration and teachers determining what assessments are given. In addition, within the field we have found that certain decisions around administration tend to go hand in hand with certain content structures. For example, we have found that assessment programs with the Full Domain content structure tend to have fixed windows, with at least the first and last assessment being required.

Depending on the aggregation approach used, the structure of administration may be more or less standardized. More generally, the state will need to determine which assessments are required for aggregation to inform the summative score. If highly standardized, the state will need to check for acceptable trade-offs for instructional purposes. For example, a state might say that three assessment modules are required to be administered within specified windows in the first, second, and third quarters of the school year. Would a district that wishes to administer more or fewer interim assessments be allowed to do so, or would that disrupt the use for summative purposes?

Implementation Issues

In addition to design decisions that shape student, teacher and leader experience, there are a number of issues related to implementation that will need to be addressed in order to successfully administer a through-year program. To be clear, there are emerging examples of through-year assessments that have been successfully administered, for example the Louisiana Innovative Assessment Demonstration authority reported operationally for the first time for the 2021-2022 school year and the Instructionally Embedded option of Dynamic Learning Maps has been reporting operationally since 2014-2015. In addition, a number of states will be reporting operationally for the first time for the 2022-2023 school year.

- Sufficient data – In a state assessment, provisions are made to support complete student testing, both so every student can complete the required tests and so every student for whom the school is accountable has fair opportunities to participate in the assessment. This involves providing, for example, extended testing windows and retesting opportunities. When there are administrations of multiple tests during the year, there may be requirements for multiple extended windows and retesting. Even then, there may be students who have missing data. The state will need to decide what to do about missing or incomplete data needed for aggregation to create the summative score or determination.

- Security – State summative assessments are typically administered under highly specified and standardized conditions to support the validity and comparability of results. This standardization includes security to ensure that student performances support the intended claims, (e.g., that the student received only permissible help and consulted only appropriate resources, and that intellectual property is highly protected (i.e., not copied). However, interim assessments often are administered under much less strict security and standardization. The state will need to determine what level of security is needed to support the summative purposes, and whether that security seriously undermines the instructional utility of the through-year assessments.

- Accommodations and supports – State summative assessments are required to provide adequate accommodations and supports to support the validity of intended claims regarding student performance and ability by reducing construct-irrelevant variance. The state will need to determine whether the constructs and construct-irrelevant variance are the same for summative and instructional purposes, and if so, then ensure that the same accommodations and supports are provided across both assessments. Often commercial interim assessments provide fewer accommodations than do state summative assessments, and accommodated performance has not been incorporated into scaling and other technical work.

- Data governance and access – In a state assessment, the state typically “owns” and has full access to the data. Districts have typically “owned” the data from instructional assessments. For through-year assessments that serve both purposes, policies will need to be agreed to that govern access to data to ensure each entity has appropriate access and that appropriate confidentiality and privacy are maintained.

Aggregation

A defining requirement, albeit a contested one, of a through-year assessment program is that information from the multiple modules will be combined to yield a summative score or determination. This remains a challenge for most current designers of through-year assessments in that few have fully specified or made operational their information aggregation procedures, let alone demonstrated the technical adequacy of the properties of the scores, designations, or other information newly provided by through-year assessments.

There are at least three sources of technical challenges to aggregation of information in most through-year assessments:

- Different purposes—e.g., summative and instruction—are optimized differently. Combining a single assessment to fulfill two very different purposes creates technical challenges
- Multiple assessments administered multiple times for data collection require combining things that are unlike in some ways. The more unlike they are, the greater the technical challenges to combining them using current measurement models and procedures. For example, an assessment scale is typically developed in reference to a specific time of performance (e.g., end of year).
- Multiple governance—particularly reflecting traditional state control over summative assessments and district/school control over interim assessments, especially for instructional purposes often results in policy tensions that are a source of non-uniformity of assessments, and of challenges to creating processes for resolving the tensions and deciding on technical solutions.

Score aggregation is as much an exercise in determining what is valued by stakeholders as it is an exercise in measurement. That is, defining the summative claim, and what is valued about student performance across the year, guides the selection of methods to produce an aggregate score. In particular one might ask whether, and if so how should:

- Performance *within-year* be valued?
- Performance *at the end* of the year be valued?
- Changes in *performance* across the year be valued?

Score aggregation is as much an exercise in determining what is valued by stakeholders as it is an exercise in measurement.

The summative claim and resulting aggregation method also need to take into account how the content domain is structured across the modules, as well as the implications that structuring has for opportunity to learn. For example, under the full domain model, resulting scores from each module (e.g., scale scores based on item response theory) can be subjected to a number of possible operations, e.g.,:

- End-of-year Only Score (or with other unit based assessments as a prior)
- Weighted Average Score
- Best Of Score (i.e., maximum score)
- Change Score (across windows)
- Weighted Change + Status
- Conjunctive (i.e., all tests must met some threshold)
- Rule Based (e.g., using either the last score or a weighted composite, whichever is higher)

Each of these methods come with their own complications, and not the least of which are stakeholder perceptions, addressing change across windows, and measurement precision.

Other models that allocate the content domain across multiple modules must aggregate the results to fully represent the domain, assuming there are unique portions of the content domain that are only assessed on one module. One approach to score aggregation in this context, which puts roughly equal emphasis on each module, assuming each module is roughly the same length, is to treat the full set of modules as if they are a single test, then create a summative score accordingly. Alternatively, scores from each module itself can be aggregated using a variety of approaches, from simple sums (e.g., if each module is a mastered standard, then the summative score could be the number of mastered standards) to more complex models (e.g., rule based aggregation methods, for example, reaching a specific level of performance on each module).

4 Emerging Examples of Through-Year Assessment Program Design

The multiple purposes of through-year assessments are often in tension, and trade-offs are required because it is not possible to optimize both purposes in a single through-year design. For example, an optimum summative test might be a certain length, and an optimum instructionally informative test might be a certain length, but the combined length of the two tests may be unacceptably long. If the state has an agreed-upon set of principles and goals clearly articulated and documented, then the decisions and rationales about each trade-off are more likely to be coherent.

Through-year assessments may become less coherent when there is "mission creep" or a trade-off was not fully specified at the design phase, but when decisions about specifics are made, they are made individually without reference to the design of the whole assessment system. Single purpose assessments may also suffer from lack of coherence, but the threats are greater for through-year assessments because they are more complex.

Keeping these tensions in mind, this section presents some prominent designs for through-year assessment. For each design, how key technical issues are addressed will be shown, and the technical strengths and challenges of each design will be briefly discussed.

Grading Model

- Multiple assessments: e.g., Four assessments modules at end of each quarter
- Combined for summative determination: Scores are combined into a weighted average

This is the simplest and perhaps most familiar model—it is quite similar to how a teacher might create a final grade. It has considerable flexibility in how the content is organized by assessments throughout the year. It has considerable flexibility in how the information across assessments is combined into a final summative determination—think of the many ways a teacher might assign points and weights to assignments throughout the year.

One key issue to a grading model is whether the purpose is to provide primarily a summative grade at the end of the year, or to provide during the year as well. The dilemma can be shown in terms of how to treat student performance that changes during the year. This is a clear expression of the challenge to decide upon the rules for score aggregation that were discussed earlier.

Adaptive for End of Year Model

- Multiple assessments: Three within-year modules and an end-of-year modules
- Combined for summative determination: Performance on within-year yield an estimate of student ability that is used to inform where the end of year summative computer adaptive test starts

This model embodies the decision to downweight student performance during the year and rely mostly on student performance at the end of the year in order to create the summative determination for the end of the year. This is shown in the fact that the student's performance at the end of the year can override the evidence collected during the year. Another advantage of this model is that the instructional and summative assessments do not have to be as tightly integrated in content or scaling since the end of year assessment is the primary source of evidence regarding the summative determination. For example, the instructional assessments administered prior to the end of the year might be commercial interim assessments, and the summative might be more like a traditional state summative assessment.

Mastery Sequence Model

- Multiple assessments: Many assessments targeted at knowledge or skills ordered according to some sequence
- Combined for summative determination: The most advanced level in the sequence where the student demonstrates mastery is the student's proficiency determination.

This model strives to take advantage of known sequences of learning. If it is established that a student must know some knowledge or have some skill at some level in order to learn the next content, then the assessment can a) focus on those competencies, and b) aggregation is simplified because it can be assumed that if a student has mastered some level, the student is also competent in the prerequisite content/levels.

An essential aspect to this model is that the learning sequences and competencies must be established. There are few "learning progressions" that have been empirically supported to date, and most have been subject to instruction at least in part, and not a psychologically inherent sequence regardless of curriculum. Thus a state through-year assessment would need to get agreement about the sequence of content, e.g., curriculum framework at least.

Supplemented Summative Subscores Model

- Multiple assessments: within-year assessment modules and an end-of-year assessment module.
- Combined for summative determination: Summative determination from the end-of-year assessment module only.
- Combined for another purpose: within-year and end-of-year assessment modules combined to provide better subscores (e.g., more, more detailed, more reliable, more contextualized) than the summative assessment could provide on its own.

This model is different in that it is designed to provide more fine-grained information at the end of the year, which could be used to inform instructional decisions—particularly programmatic improvement—as well as whatever within-year information is available from the assessment module. This is a way to reduce the footprint of the summative assessment while providing useful information at the end of the year.

- Some of the types of information that might be provided include:
- A projected proficiency on the end of year summative assessment from each administration of the within-year assessment modules.
- Within-year assessment modules that are on the end-of-year assessment scale, allowing direct scale comparisons between modules, such as for within-year growth on the within-year assessment module scale.
- Subscores related to the end-of-year assessment module's constructs and reporting categories that are more reliable than could be accomplished with the summative items alone.
- Subscores related to the within-year assessment modules constructs and reporting categories to permit comparison with other instructionally oriented information reported during the year.

Note that it may be difficult to provide any one of these types of information, and impossible to provide all in the same through-year assessment.

5 Evaluating the Technical Quality of Through-Year Assessment Designs and Assessments

A state or other sponsor or user of through-year assessments will need to evaluate both the design and the particular assessments.

Evaluation of the assessment quality should follow the professional guidelines established in the *Standards for Educational and Psychological Testing* (2014), jointly sponsored by AERA, NCME, and APA. These professional organizations establish the guidelines for assessments in use in a very wide range of circumstances in the United States, including educational K-12 settings. Although more complex because a through-year assessment consists of multiple assessments that are combined in some way for the summative purpose, the *Standards* still apply. The *Standards* also apply to assessments that are used for instructional purposes, especially when those assessments are sold commercially or sponsored by an organization that intends the usage to be credible and impactful. It is true that the current *Standards* say classroom assessments developed by teachers are not expected to document compliance with the *Standards*, perhaps because individual teachers are not be expected to devote the resources necessary to document the quality of their assessments, especially when they are limited to transitory, personal use. Certainly assessments adopted by states for through-year use should be evaluated in regards to the *Standards* for validity, reliability, and fairness. An additional requirement for assessments being used to meet federal assessment and accountability purposes are also required to undergo federal Peer Review.

Evaluation of the assessment quality should follow the professional guidelines established in the *Standards for Educational and Psychological Testing* (2014), jointly sponsored by AERA, NCME, and APA.

Of course, through-year assessments differ in some significant ways from traditional end-of-year summative assessments, and so even for federal Peer Review there will need to be some adjustments and additions. Through-year assessment are just recently being submitted for Peer Review, and there may be considerable variation in the models, so these comments are "informed guesses" rather than a summary of what has actually be required of through-year assessment. The specific through-year assessment design will dictate those modifications. In a Gradebook model where all the within-year assessment modules contribute substantially to the summative determination, we expect that all the within-year assessment modules would be subject to full Peer Review. In contrast, in an Adaptive for End-of-Year model, where the within-year assessment modules are optional because the summative determination depends (almost) exclusively on the end-of-year assessment, Peer Review might concentrate only on the end-of-year assessment module.

Where the whole set of assessments in the through-year program are subject to Peer Review, we expect that alignment of assessments to content standards would be examined closely, especially where the instructional uses of the through-year assessment encourage off-grade assessment and/or partitioning of the full set of state content standards across multiple assessments. Showing that adaptive tests meet federal requirements of all students being assessed on the state content standards has required careful construction of evidence different than an alignment study for fixed forms; this could be more complex with adaptivity in play across multiple assessments, especially with dependencies between assessments, and with the possibilities of missing data if a student did not complete all the through-year assessments.

Comparability is assumed for federal uses–results must be able to be aggregated across students and over time for many purposes. As discussed previously, through-year assessments by their nature often pose greater challenges to comparability–they are trying to serve multiple purposes, they consist of multiple assessments, they are administered at multiple times, and they may be administered under varying administrative conditions. Each of these aspects must be addressed in Peer Review to show how there is sufficient comparability to allow fair, accurate, and reliable interpretations and uses, such as for accountability identifications of consistently underperforming student groups.

Evaluating most uses of education assessments involve a type of program evaluation, which is informed by the *Program Evaluation Standards* more than the *Standards for Educational and Psychological Testing*. An *Interpretive/Use Argument* is largely focused on the degree to which the interpretation can be supported, although it does touch on use. Use and effects are specified in the *theory of action* or *logic model* for an assessment. Evaluation of the theory of action or logic model involves program evaluation in logic and substance, because the focus is not on whether the interpretations are valid, reliable, and fair, but on to what degree the uses meet standards of utility, feasibility, propriety, and accuracy (d'Brot, 2022). This is especially true when the assessment results are not the "treatment," or not the primary aspect of the actions intended to make a difference in student learning (e.g., assessment informs what instructional actions to take, but program evaluation would examine whether the assessment provided accurate information, and whether the instruction led to increased student learning). Both summative and instructional uses of assessment are intended uses and impacts, and so both aspects should be subject to program evaluation.

Initial supports are provided in this document for making three types of evaluation regarding through-year assessments:

- Policy evaluation for whether through-year assessments are likely to meet the state's purposes and constraints
- Technical quality evaluation of the through-year assessment system in terms of interpretation
- Program evaluation of the use and impact of the through-year assessment

Summary and View to the Future

Through-year assessment programs intentionally bring state educational agencies closer to districts, schools and classrooms, and based on the current dialogue, closer to classroom instruction. Doing so means dealing with a number of technical challenges in relation to tensions between supporting desired instructional use-cases while simultaneously providing an annual determination that supports ESSA compliant systems of school identification and support.

These tensions are not limited to technical concerns; indeed, there is great potential for unintended consequences, including consequences in which the intended instructional use-cases are jeopardized due to their connection to accountability, that the summative score is perceived as unfair due to incorporation of within-year information, or the perception of the administration of state assessments during the year is an over-reach by state educational agencies. Ultimately, the success of any through-year program is an empirical question, and one that is highly contextual. A design that might work well for one state may not work well for another state, so it is important that each state educational agency considering a through-year program consider how to carefully thread the needle between the tensions inherent in any through-year program.

As we note previously (see Dadey & Gong, 2017), through-year programs require a great deal of investment, well above and beyond a typical state summative program. Through-year programs also require much greater tailoring to both local and state context than typical summative assessment programs. States can and do adopt summative assessment programs that are "off the shelf" and implement them successfully.

We argue that such an approach for a through-year program is not likely to meet that program's intended goals, and instead the success of a through-year program is highly contextual, and requires careful tailoring to state, and local, contexts. Even with such tailoring, the success of any through-year program is not guaranteed, as the program must navigate a complex number of issues, including balancing instructional and summative uses (and in doing so, address the possibility that the additional purposes, including those related to instruction, may be compromised due to pressures related to the use of the through-year assessment results to create summative determinations, i.e., Campbell, 1976).

We hope this paper will support the field in effectively wrestling with the complex and interconnected issues inherent in the design of through-year programs. We look forward to emerging work that will clarify choices, provide technical solutions, and inform policy decisions in support of better assessment systems that support increased student learning and more equitable schooling.

References

Abrams, L. M., McMillan, J. H., Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Education Assessment, Evaluation, and Accountability*, 27, 347-375.

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

Provost, L., & Bennett, B. (2015). What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress*, 36-43.

Bennett, R. E., Kane, M. and Bridgeman, B. (2011). Theory of action and validity argument in the context of through-course summative assessment. Paper presented at the 2011 Invitational Research Symposium on Through-Course Summative Assessments: Atlanta, GA. Retrieved from: https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Bennett_Kane_Bridgeman.pdf

Campbell, Donald T (1979). "Assessing the impact of planned social change". *Evaluation and Program Planning, 2*(1), 67–90.

Clark, A. K., & Karvonen, M. (2021). Instructionally embedded assessment: Theory of action for an innovative system. Frontiers in Education, 6. Retrieved from: https://doi.org/10.3389/feduc.2021.724938

Cole, S. K., & Swanson, C. (2022). Content Progressions & Clustering Across Instructional Materials: Viability for Supporting the Design of a Through-Year Assessment Model. Smarter Balanced. Retrieved from https://portal.smarterbalanced.org/library/en/sb_content-progressions_through-year-assessment.pdf

Crane, E. W. (2010). Building an Interim Assessment System: A Workbook for School Districts. Council of Chief State School Officers. Retrieved from: https://www.wested.org/resources/building-an-interim-assessment-system-a-workbook-for-school-districts

Dadey, N. (2019). Providing Flexibility to Schools and Districts while meeting ESSA Assessment Requirements [Blog]. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://www.nciea.org/blog/considering-interim-assessments-and-summative-information

Dadey, N. (2018). When It Comes to Getting Summative Information from Interim Assessments, You Can't Have Your Cake and Eat It Too [Blog]. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/blog/when-it-comes-to-getting-summative-information-from-interim-assessments-you-cant-have-your-cake-and-eat-it-too

Dadey, N. & Badrinarayan, A. (2022). In Search of the "Just Right" Connection Between Curriculum and Assessment [Blog]. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://www.nciea.org/blog/in-search-of-the-just-right-connection-between-curriculum-and-assessment

Dadey, N., & Gong, B. (2017). Using interim assessments in place of summative assessments? Consideration of an ESSA option. Washington, DC: Council of Chief State School Officers. Retrieved from: https://ccsso.org/resource-library/using-interim-assessments-place-summative-assessments-consideration-essa-option

D'Brot, J. (2022). A 10-step guide for incorporating program evaluation in education into the design of educational interventions [Blog]. Dover, NH: National Center for the Improvement of Educational Assessment.

Frechtling, J.A. (2007). *Logic modeling methods in program evaluation*. San Francisco, CA: Jossey-Bass.

Georgia Department of Education (2018). Georgia's application for the innovative assessment demonstration authority under section 1204 of the Elementary and Secondary Education Act (ESEA). Atlanta, Georgia: Author. Retrieved from: https://www2.ed.gov/admins/lead/account/iada/gaiadaappdec2018.pdf

Gong, B. (2010). *Using Balanced Assessment Systems to Improve Student Learning and School Capacity: An Introduction*. Dover, NH: National Center for the Improvement of Educational Assessment.

Gianopulos, G. (2019). From Through-Course Summative to Adaptive Through-Year Models for Large-scale Assessment: A Literature Review. NWEA Research. Retrieved from: https://www.nwea.org/uploads/2021/06/From-Through-Course-Summative-to-Adaptive-Through-Year-Models-for-Large-scale-Assessment_NWEA_literatureReview.pdf

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (17-63). Westport, CT: Praeger.

Marion, S., & Leather, P. (2015). Assessment and Accountability to Support Meaningful Learning. *Education Policy Analysis Archives, 23*(9). Available Online: https://epaa.asu.edu/ojs/article/view/1984

National Center for the Improvement of Educational Assessment. (2021). (8) Through-Year Assessment Virtual Convening. Dover, NH: Author. Retrieved from https://www.nciea.org/library/technical-logistical-issues

- Session 1: Dadey, N., Gong, B., Lorié, W., & Marion, S. (2021). Through-Year Assessment Virtual Convening: Session 1: Definition, Aims, and Use Cases. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/library/definitions-aims-and-use-cases
- Session 2: Gong, B., Dadey, N., Lorié, W., & Marion, S. (2021). Through-Year Assessment Virtual Convening: Session 2: Definition, Aims, and Use Cases. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/library/claims-designs-and-evidence
- Session 3: Lorié, W., Dadey, N., Gong, B., & Marion, S. (2021). Session 3: Technical & Logistical Issues. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/wp-content/uploads/2022/07/Session-3_TechnicalLogisticalIssues_Presentation.pdf

- Session 4: Marion, S., Lorié, W., Dadey, N., & Gong, B., (2021). Through-Year Assessment Virtual Convening: Session 4: Session 4: Threading the Needle . Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/library/claims-designs-and-evidence/

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).

Pellegrino, J. W.; Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10019.

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

Salvia, J., Ysseldyke, J., & Bolt, S. E. (2012). *Assessment: In special and inclusive education*. Cengage Learning.

U.S. Department of Education. (2018). A state's guide to the U.S. Department of Education's assessment Peer Review process. Washington, DC: Author. Retrieved From: https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf

Appendix A

History of Through-Year Assessment

Statewide summative assessment is, and has been, under continual pressure to change. One way this pressure has manifested is in a push for "through-year" or "through-course" assessment. "Through-course assessment" gained prominence as an assessment design after being defined within the call for the Race To The Top grant program (U.S. Department of Education, 2010, p.18178) and encouraged by the now closed Enhanced Assessment Grants Program (U.S. Department of Education, 2016). As a result of these funding streams, through-year assessment was became the subject of much attention and research during the early development and implementation of the general consortia assessments, Smarter Balanced and the Partnership for Assessment of Readiness for College and Careers (see the series of papers released in 2011 as part of an Invitational Research Symposium on Through-Course Assessment; Bennett, Kane & Bridgeman, 2011; Ho, 2011; Sabatini, Bennett & Deane, 2011; Wise, 2011; Zwick & Mislevy, 2011). In this line of work, through-year assessment programs were generally referred to as "through-course" assessment[4]. Much of the emphasis on this line of work was on incorporating performance assessment tasks throughout the year, whereas current efforts are more expansive, aimed at addressing a number of different problems of practice. We suggest that "through-year" be used as an umbrella term to encompass a number of prior terms, including through-course and "statewide interim assessments" (e.g., as specified under ESSA). We do so, in part, because the term through-course has, at times, been used to indicate that the assessment program is tied to a specific course of study, whereas through-year is more general. This distinction is not made uniformly or consistently, so here we adopt the term through-year.

Following this initial burst of interest in through-course or through-year assessment during the beginning stages of the general assessment consortia, interest waned as the proposed models were seen as potentially constraining to curriculum and instruction as well as difficult to administer due to the additional time needed for the multiple parts of the assessment. However, at least one alternate assessment consortia, namely Dynamic Learning Maps, has continued as a through-year assessment program, offering "Instructionally Embedded" and "Year End" models since 2016 (see also Clark & Karvonen, 2021).

We suggest that "through-year" be used as an umbrella term to encompass a number of prior terms, including through-course and "statewide interim assessments" (e.g., as specified under ESSA).

[4] One might also argue that the focus of through-course was different, at least in part. Much of the goal of through-course work from this period was to provide greater content coverage, whereas through-year design aims to serve a wider array of uses.

Occurring during the same time period, states began to provide and support interim assessments (Perie, Marion & Gong, 2009) alongside their statewide summative assessment. In 2019, Dadey & Diggs (2019a) found that twenty-five states provided some form of interim assessment based on a scan of all fifty state websites, signaling a shift towards greater interest in practices based on interim assessments and potentially an early sign of a shift towards through-year assessment, as state supported voluntary interim lays a foundation for a shift to statewide through-year assessment. However, some states that have voluntary state supported interim assessments have intentionally chosen not to develop through-year programs due to the entailed challenges. Ultimately, all of these efforts suggest that state educational agencies across the nation are deeply interested in greater coherence in assessment practice.

More recently, a provision of the Every Student Succeeds Act (ESSA) of 2015 outlines an approach to through-year assessment, albeit through with different language, stating that statewide assessment 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); see also Dadey & Gong, 2017). ESSA also contained several options for assessment flexibility, including the Innovative Assessment Demonstration Authority (IADA) waiver, which allows states to develop and use a new assessment with a subset of schools and districts, while still meeting federal requirements for statewide annual testing. Three states, Georgia, Louisiana and North Carolina pursued and where granted flexibility under IADA in late 2018 and 2019 to develop what we now refer to as through-year assessment programs. These three states have maintained their current statewide systems while also developing through-year assessments programs, ideally allowing the space for innovation. Currently, only Louisiana has reported results operationally from their IADA program in Spring 2022. North Carolina intends to report operationally in Spring of 2023.

Three states, Georgia, Louisiana and North Carolina pursued and where granted flexibility under IADA in late 2018 and 2019 to develop what we know refer to as through-year assessment programs.

Trailing slightly behind the under the IADA waiver were efforts in two states, Alaska and Nebraska, to develop through-year programs as the statewide assessment. These efforts involved procuring new assessment programs in place of their previous statewide assessments. These two programs, AK STAR: Alaska System of Academic Readiness and NSCAS Growth: Nebraska Student-Centered Assessment System Growth, are built around the already exiting NWEA MAP Growth interim assessment. Both of these programs have been under development since the 2020-2021 school year, if not earlier and plan to report operationally in Spring of 2023. Maine, whose development started later in 2021-2022, also plans to report operationally in Spring 2023. This means that in late Spring 2023 there will be five states who will have reported operationally using a through-year approach: two under IADA, Louisiana and Nebraska, and three statewide Alaska, Maine and Nebraska.

Overlapping with these efforts is an explosion of interest in through-year assessment as the field has, partially, appeared to emerge from the pandemic. In late 2021, at least ten states were pursing through year models[5] (Dadey, Gong, Lorié & Marion, 2021), with the number standing at thirteen in late 2022 (Ed First, 2022). This explosion is likely due to a number of interrelated factors. First and foremost, the field – not the least of which includes students, parents, teachers and leaders – have critiqued statewide summative assessment as providing too little utility while having too great a footprint. Thus the field is eager for change. This eagerness may be interacting with the "post" pandemic period in which many seem to be willing to experiment or change previously accepted approaches and practices like statewide summative assessment.

In addition to this general orientation of the field, some vendors are acting as market disruptors, philanthropic organizations and the US Department of education have provided funding and state legislators have created legislative requirements. In terms of market disruption, a number of states are working with vendors who are acting as market disruptors, as these vendors are new to the state summative assessment landscape. Likely, such market disruption will continue, as a number of new vendors look to develop through-year programs and current vendors react to the shifting landscape. Much, although not all, of this disruption has been supported or sustained by grants provided by educational philanthropies. These philanthropies have provided multiple waves of funding for investigation and development of through-year assessment program development[6]. The US department of education, through legislatively required cycles of the Competitive State Assessment Grants has also provided funding to support through-year programs. The most recent round of awards in 2022 included grants to several states, including Louisiana, Montana, and Missouri, whose proposals focused on the development of through-year assessment programs. Finally, at least two states – Florida and Texas - have had legislation essentially require a through-year approach.

[5] This count does not include states that are currently using the DLM Institutionally Embedded options. Also note that the New Hampshire PACE program, not included here, could also be considered a through-year program.

[6] See, for example, the line of work funded by the Bill & Melinda Gates Foundation, Walton Family Foundation, and the Chan Zuckerberg Initiative and organized by Ed First.

Appendix B

Purposes with Additional Detail and Implications for Design

Purpose	Implications for Design	Note
Save time	Use existing commercial interim assessments in lieu of the state summative assessment;	No state that has submitted for Peer Review as tried this; no vender we are aware of advises this is their preferred approach to through-year assessment
Provide greater instructionally relevant information than current summative test provides	A through-year assessment design might fulfill this purpose	State should make clear the cost/benefit of a through-year versus a balanced assessment approach, where the latter does not require the assessment modules to contribute to the summative determination
Maintain current district flexibility regarding choice of assessments	Allow district choice of different commercial assessments	Not permitted under ESSA except at the high school level ("nationally recognized"); the ESSA provision allowing for multiple "modular" (interim) assessments requires the state to identify a single assessment
Maintain current district flexibility regarding administration of interim assessments	Allow district and/or educator choice regarding time during year, in relation to instruction, and administration conditions	When used for accountability, will the state require more standardization in test windows? Will educators feel they have to administer after instruction is "complete" to get highest score? Will the state require strict security and administration standardization procedures as it does for its current summative assessment?
Assess aligned content/skills	Assessment taps valued learning targets	For summative, usually assesses (almost) all of the grade-level content standards For instructional, may assess many fewer in any one assessment, and some may not be included in the state's grade-level content standards, or may be off-grade. The through-year design will need to reconcile these possible differences.