DESIGNING AND EVALUATING INNOVATIVE STATE ASSESSMENT PROGRAMS: A FRAMEWORK FOR STATE EDUCATION AGENCIES

May 2024

W. Christopher Brandt, Ph.D.
Nathan Dadey, Ph.D.
Carla Evans, Ph.D.

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

TABLE OF CONTENTS

Acknowledgments:
We acknowledge the support and feedback from Juan D'Brot and Scott Marion at the Center for Assessment, whose ideas and suggestions improved this paper. Thanks also to Catherine Gewertz, whose edits improved the clarity of our work.

Brandt, W.C., Dadey, N., and Evans, C. (2024). *Evaluating innovative state assessment programs: A framework for state education agencies.* National Center for the Improvement of Educational Assessment.

INTRODUCTION

The past several years have produced a surge in interest in improving state assessment programs. More than a dozen states are in various stages of rolling out new innovative state assessment programs statewide (Dadey & Gong, 2023; Education First, 2023). Here we use "innovation" broadly, to describe programs that are being designed to address real or perceived problems with the typical domain-sampled, end-of-year summative assessment that has characterized state testing since the passage of the No Child Left Behind Act of 2001. Most states designing innovative assessment programs are doing so as part of their regular program development. A minority are developing their new assessment programs under the Innovative Assessment Demonstration Authority (IADA) of the Every Student Succeeds Act (ESEA, 2015).[1]

Much of this innovation in state assessment is aimed at addressing longstanding areas of unhappiness with state assessments: states want to streamline assessment activities, enhance the instructional utility of assessment results, and improve assessment experiences for students. Designing and redesigning state assessment to meet these goals often involves working quickly to meet legislated timelines or legislative turnover. And working quickly often means that good practices—particularly those related to iterative design and program evaluation—can be overlooked. Without a systematic and robust process of designing and evaluating these programs, educators and policymakers risk losing critical insights into what works well, what's going wrong, and why.

This paper presents a framework that state education leaders can use to design and evaluate an innovative assessment program. In doing so, it is meant to orient and frame design and evaluation. Although this paper focuses on innovative state assessment programs, the framework can be used to guide the design and evaluation of any state assessment program. It does not describe, in detail, the step-by-step implementation of an innovative system, as doing so requires far more time and space than provided within a single paper. However, by framing the process, we hope that state staff and their partners can easily identify *where* they need to dig deeper into available resources.

THE NEED FOR A ROBUST DESIGN AND EVALUATION APPROACH

Currently, far too many state assessment programs—innovative or not—are being undertaken without iterative design and program evaluation. It is easy to see why this is the case. Simply developing and implementing an innovative assessment system is a sizable task. If an educational agency and its partners are at or beyond capacity by just building a program, how could they possibly find time and resources to iterate and evaluate? We argue that the opportunity cost is too great *not* to.

There are two key pitfalls that often occur in these kinds of contexts. First, without iterative design, an innovative state assessment program can get "stuck," with initial exploratory designs becoming the final solution, when often additional adjustments could yield better outcomes. This is, in part, because stages of design and evaluation, like those we propose below, are often not built into the initial program design. Second, without evaluation, whether a program works as intended is unclear. Well-designed and executed evaluations produce a wealth of historical knowledge that informs future directions in assessment. Moreover, evidence from such evaluations inform better ideas that improve learning for the next generation of students.

[1] The IADA allows states to administer an innovative assessment in place of the statewide assessment for a subset of schools, subject to certain constraints. This means that states, under the IADA, can pilot a new program while running their current one, *without* the need for double testing.

The framework we present is an attempt to mitigate these pitfalls. This framework also presents an idealized version of design and evaluation; most programs will never achieve the kind of robust, multistep evaluation proposed. In many cases, they don't need to. Instead, what is needed is a careful application of the thinking contained within the framework to identify what kinds of design and evaluation processes need to be implemented.

Prioritization is key. What evaluative questions are most important to answer at each phase of the design process? What questions can be addressed with the limited resources available? Limited resources often force difficult choices, and iteration and evaluation are often the first things to go. But we argue that the value of even a partial iteration and evaluation outweighs the drawback of not knowing why innovations work or, perhaps more probably, why they did not work and how to improve them. Moreover, there is value in considering each of the four phases within the framework, even if the program doesn't progress through that phase explicitly. Often, program development is so fast that one or more phases get overlooked, leading to designs that might not live up to their intended goals and purposes.

The need for well-designed and executed evaluations exists over and above any specific regulatory requirements for evaluation. For example, states may consider the federal peer review process to substitute for evaluation; however, peer review and evaluation serve two fundamentally different purposes. Peer review exists to ensure that statewide assessments are fair and secure, tests are reliable and air, and test score interpretations are valid. At its core, evaluation is about learning how to improve. A well designed and executed evaluation informs:

- sound judgments about whether a program is achieving its goals,
- evidence-based determinations about how and for whom the program is working and not working, and
- changes that may be introduced to improve outcomes for learners.

The need for well-designed and executed evaluations exists over and above any specific regulatory requirements for evaluation. For example, states may consider the federal peer review process to substitute for evaluation; however, peer review and evaluation serve two fundamentally different purposes.

This framework is designed to accomplish the three objectives listed above. The framework provides guidance for state and local agencies to build evidence of a program's effectiveness by integrating evaluation into the process of developing and taking a program to scale. Additionally, the framework draws on principles of improvement science to improve program components through small-scale plan-do-study-act trials.

EVALUATION AS A FUNDAMENTAL COMPONENT OF CONTINUOUS IMPROVEMENT

Evaluation and continuous improvement share several overlapping principles and processes. Continuous improvement supports state and local agencies' efforts to change, improve, and evaluate programs and practices (Shakman et al., 2020). Continuous improvement recognizes that a

robust evaluation design depends on a clear and well specified program design. Models of continuous improvement refer to three essential questions, which are also useful for designing program initiatives and, in turn, informing evaluation planning (Bryk et al., 2015).

Evaluation and continuous improvement share several overlapping principles and processes.

1. What is the problem we are trying to solve?
2. What changes are being proposed to address the problem?
3. How will we know that these changes are producing the intended results?

In our experience, organizations can generally give clear and succinct answers to the first two questions.

What is the problem? There is widespread agreement among state leaders that testing has reached a tipping point. Leaders want to address a variety of problems: Students take too many tests, and few, if any, are instructionally useful. State test reports are often not available in a timely manner and test results are often misused. The need to improve assessment has sparked a wide range of promising ideas about how testing can be reduced, streamlined, and made more instructionally useful (USDE, 2016).

What changes are being proposed to address these problems? States piloting innovative state assessment programs have, for the most part, developed theories of action to address these problems. Theories of action outline an initial vision that describes (1) what resources are needed, (2) how the program will be implemented, (3) the initial underlying mechanisms, activities, and assumptions that will support program implementation, and (4) how those mechanisms and activities will produce the intended short- and long-term outcomes. A logic model is often produced from the theory of action to aid in developing a robust evaluation plan.

Theories of action and logic models provide the blueprint to design and implement robust and improvement-oriented evaluation designs. Appendix A includes select states' theories of action for innovative assessment programs. For example, Montana and several other states have designed theories of action for through-course assessment solutions. Montana plans to improve the instructional utility of test information by developing multiple smaller (modular) tests, administering them flexibly throughout the year, and then returning results to educators more quickly.

The third question, however, is often left underspecified, if it is addressed at all: **How will we know whether these new innovative state assessment programs will produce the intended results?** This question is where organizations, including state education agencies, often struggle to describe a clear path forward and evaluation plan. The U.S. Department of Education requires state awardees to submit annual performance reports (APRs) and state-sponsored evaluations of federally sponsored programs such as the Innovative Assessment Demonstration Authority (IADA) and Competitive Grants for State Assessments (CGSA). However, state departments often have limited funds to support deeper research and program evaluations to address these evaluative purposes. Moreover, programs like IADA are unfunded, making it difficult for IADA states to evaluate whether, how, and for whom, test program innovations are working. Finally, states that are developing these programs without federal support often struggle to secure staffing and funding to implement ongoing evaluation and continuous improvement.

Evaluation is an essential component in continuous improvement. Evaluation is the mechanism that allows stakeholders to know what is working as intended and leading towards intended outcomes, and what needs to be adjusted. Examples of evaluation questions include:

Evaluation is an essential component in continuous improvement. Evaluation is the mechanism that allows stakeholders to know what is working as intended and leading towards intended outcomes, and what needs to be adjusted.

- **To what extent is the program meeting its goal and working as intended?** Is the innovative assessment program working as designed? Is it producing the desired effects?
- **Where, for whom, how, and under what conditions is the program working as intended?** Even if the program does not initially work as designed, perhaps there are specific pockets or populations of students for whom the program seems to be achieving the desired outcomes. Where do we see such variation in program implementation and/or outcomes? Where or with whom is the program working as intended? Where/with whom is it not working as intended? What seems to be causing these variations to emerge? How might changes, or flexibility, in program resources or inputs improve implementation and outcomes in specific contexts and/or among specific populations of students (e.g., the most disadvantaged or underserved students? Under what conditions does the program work best, and for whom?
- **How can the program be improved to meet its stated goals?** What problems persist as the program continues to scale? How can the program respond effectively to the changing needs and priorities of end-users? How can the program take advantage of new and evolving technologies to better accomplish its purposes and goals?

We acknowledge that there are many ways of approaching program evaluation and continuous improvement. Our framework describes general principles for building evidence to guide program improvement. It stops short of recommending specific formative or summative evaluative approaches or strategies.

Our framework describes general principles for building evidence to guide program improvement. It stops short of recommending specific formative or summative evaluative approaches or strategies.

For example, iterative improvement cycles are an essential component of formative program evaluation, and there are numerous evidence-based approaches and strategies through which iterative improvement cycles can be implemented. Networked improvement communities, design-based implementation research, and implementation science are approaches to iterative improvement that include a promising evidence-base (LeMahieu et al., 2017). Similarly, summative program evaluation typically relies on experimental and quasi-experimental approaches. Both have their strengths and weaknesses. The selection and use of these particular approaches will depend on factors such as:

- The specific questions to be addressed,
- The quality and expansiveness of the state's measurement infrastructure (e.g., ability to longitudinally track students; link data across departments and agencies),

- The availability of resources, and
- The feasibility and practicality of randomly assigning schools to conditions.

Though we do not subscribe to any one approach, we encourage state agencies to do their research and select an approach that is well-researched and best suited for their context.

A FRAMEWORK FOR DESIGNING AND EVALUATING INNOVATIVE ASSESSMENT PROGRAMS

Figure 1 presents our framework for designing and evaluating a state assessment program. This framework is well suited for innovative assessment programs, but applies to any state assessment program, innovative or not. We acknowledge that the framework's components and phases represent an idealized scenario: States are rarely able to develop state assessment programs in such a linear and tiered fashion. Yet we believe the framework is useful because even if the process is messy or some phases are combined, it still represents all phases and considerations a state should consider over time.

The framework is organized into a planning phase followed by four implementation phases. The four implementation phases are adapted from "tiers of evidence" that are defined within the Every Student Succeeds Act (ESSA). These tiers were designed to guide the collection of evidence for the evaluation of educational interventions. Here, we adapt them to work with assessment programs. In brief, the five phases are:

- **Program Planning.** This phase focuses on establishing a clear vision that explains what problem(s) the program is meant to solve, the goals and use cases that align to these problems, and how the program is meant to function, typically captured through a theory of action. Additionally, plans for the implementation of the program and its evaluation are developed.
- **Program Design and Prototyping.** The design and evaluation in phase one should be designed to accomplish two goals: (1) establish evidence of a program's theoretical rationale and (2) support the iterative development of the most essential program components for implementation.
- **Program Pilot.** In phase two, the design and evaluation focuses on establishing evidence of the program's promise for achieving its goals in a targeted set of districts, grade levels and subject areas. Assessment designers should be ready to pilot the assessment program, or major program components, in a small and representative sample of schools. Schools participating in the pilot should be familiar with, and generally supportive of, the new assessment program design.
- **Program Expansion.** Phase three often consists of an efficacy trial. An efficacy trial determines whether an intervention produces the expected result *under ideal conditions.* In the context of a state assessment program, this means that participants (e.g., schools, leaders, teachers) buy into the program's promise of success and are motivated to implement the program as designed.

The framework assumes that there is no end point to innovation. Given the speed of change in the information age, evaluation and continuous improvement must become a cyclical and ongoing process that is embedded in the fabric of an organization.

- **Program at Scale.** Phase four examines the program's effectiveness at scale. In this phase, specific subjects and/or grade levels will be ready for a full-scale rollout and evaluation at different time intervals.

These phases are connected to primary research questions, design activities, formative and summative evaluation processes and the ESSA tiers of evidence.

The framework is not specific to evaluating innovative assessment programs; it can be applied by a state to evaluate a variety of programs, including the traditional state assessment program.

The framework assumes that introducing a new statewide testing program is a complex process that requires extensive coordination, input, and iteration over multiple years. The framework follows, roughly, the implementation timeline of an assessment program, from the initial conceptualization to full scale implementation.

Additionally, the framework assumes that there is no end point to innovation. Given the speed of change in the information age, evaluation and continuous improvement must become a cyclical and ongoing process that is embedded in the fabric of an organization. Finally, as noted above, the framework is not specific to evaluating innovative assessment programs; it can be applied by a state to evaluate a variety of programs, including the traditional state assessment program.

Below in Figure 1,[2] we unpack the framework by describing key steps in the evaluation process and showing the progression of changes that should occur as evidence of efficacy builds over time. We discuss the key steps in program design (shown in green), which are used to design formative and summative approaches to evaluating the program (shown in blue).

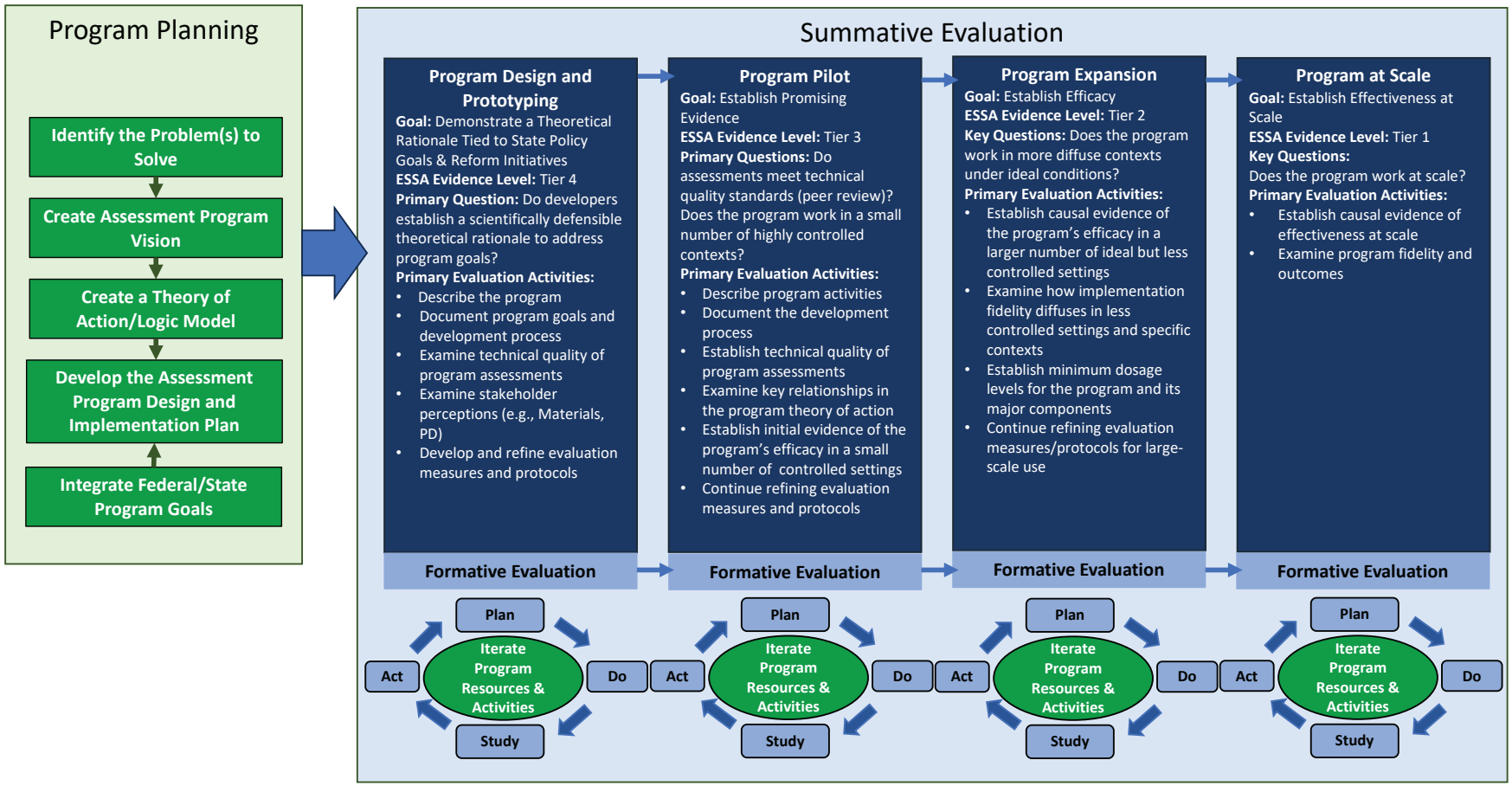
PROGRAM PLANNING PHASE

When designing for innovation, the program planning phase focuses on establishing a clear program vision and goals, along with use cases that describe how the program is meant to function for different types of users, such as district leaders, teachers, or students. The program vision is typically captured through a theory of action.

Developers must also consider how policy, structural, and practical constraints will influence program implementation. For example, as North Carolina was developing plans for a new personalized assessment program, state legislators had proposed multiple bills that, if passed, would require that the state "move toward a through-grade assessment model." Many people in education were also calling for state tests to produce more useful instructional information to guide teaching and learning. And federal law already established several constraints related to test design, administration, and use. North Carolina's design team considered these constraints as they developed a new personalized assessment program under IADA. Doing so was essential to integrating policies and priorities into program design.

[2] A complementary table is provided in Appendix A.

Figure 1: A Framework for Evaluating Innovative Assessment Programs



These steps and considerations are described in more detail below.

Identify the Problem(s) to Solve

Innovation should start with a set of well-defined problems that it is meant to solve. As indicated above, the problem evolves within a particular federal and state context, and potential solutions to address the problem must comply with federal and state laws for annual testing and accountability uses. Common problems that state assessment programs attempt to address include:

- Overlap and redundancy across state and local assessment programs,
- Too much statewide summative test preparation and testing, and
- Existing assessments' inability to provide meaningful instructional information for classroom teachers.

Create the Assessment Program Vision

The development of an innovative assessment program should be guided by a clear vision that explicates the intended purposes and uses of an assessment program and how various components of the program work together to ultimately improve student learning outcomes.

The development of an innovative assessment program should be guided by a clear vision that explicates the intended purposes and uses of an assessment program and how various components of the program work together to ultimately improve student learning outcomes.

Many external factors collectively influence interested parties' vision for the innovative assessment program, perceptions of assessment problems, and potential solutions. Examples of external factors include federal and state policies, a state's resources and capacity for reform, and as well as state and local culture and norms. For example, the federal IADA program places technical constraints around the design of the innovative state assessment program and how the information is included in school accountability. States also have varying levels of resources and staffing capacities that affect a state's ability to hire a third-party evaluator or conduct the evaluation in-house, especially in the midst of competing demands and priorities.

Clarify the Goals of the Innovative Assessment Program

The assessment problem(s) should inform the development of specific assessment program goals and use cases (i.e., the purposes and uses for which each assessment in the program is designed to address). These goals must be specified in detail in a state's theory of action. Building on the problems noted above, these goals must include:

- Aligning and streamlining assessment programs for monitoring student performance and informing school accountability, or
- Improving the instructional utility of interim and annual summative testing programs.

Create a Theory of Action

The state's policy priorities, goals, values, and intended outcomes inform the assessment program's theory of action.[3] A theory of action is a set of hypotheses, or assumptions, that describe how a

[3] The Center for Assessment has several resources to support the development of a theory of action, which are available in our resource library. For example, see: https://www.nciea.org/library/developing-a-theory-of-action-for-your-balanced-assessment-system-how-to-develop-one-and-what-to-do-with-it/. See also: https://www.nciea.org/library/a-theory-of-action-to-guide-the-design-and-evaluation-of-states-innovative-assessment-and-accountability-system-pilots/. For example, see https://www.nciea.org/wp-content/uploads/2021/01/DG-ToA-BAS-SCASS-Jan-2021C-1.pdf

given program, as designed, will support the achievement of specified goals. A theory of action is often proposed as a series of "if-then" statements. The if-then statements break down the assessment program's development process into its component parts, and illustrate how the developed program will be implemented, and how it will achieve the state's vision at scale. The assessment program's theory of action informs a subsequent logic model and evaluation design. Appendix B includes several theories of action that illustrate how states are addressing common problems of practice and achieving innovative assessment goals.[4]

A theory of action is a set of hypotheses, or assumptions, that describe how a given program, as designed, will support the achievement of specified goals.

Theories of action and logic models represent overlapping ideas, and they sometimes produce similar documentation to support an evaluation. Program designers tend to develop the theory of action. Evaluators translate the theory of action into a logic model. In a logic model, each program component—from item and test development to professional development, test reporting, interpretation, and use of results—is carefully mapped out to show the causal chain of events that lead to intended goals and outcomes.

The logic model makes the causal chain of events explicit (if the theory of action has not already done this) by describing and connecting the necessary program resources to specific activities (e.g., development processes, teacher training), and then connecting those activities to their requisite outputs (e.g., improved instruction) and outcomes (e.g., improved student engagement and learning).

Logic models are important resources for evaluators because the added specificity allows the evaluator to connect data collection and analysis methods to each phase.

Logic models are important resources for evaluators because the added specificity allows the evaluator to connect data collection and analysis methods to each phase. For example, evaluators may use a logic model to inform the development of checklists to ensure that schools have access to necessary resources. Similarly, the description of core activities and outputs informs evaluators' development of interview protocols, observation protocols, surveys, and extant data (e.g., test scores) needed to establish evidence of program fidelity and program effectiveness. In this way, the logic model serves as an evaluation blueprint to examine the implementation process and investigate for whom, how, and under what conditions a program works as designed.

Develop the Assessment Program Design and Implementation Plan

The program's theory of action and logic model acts as a roadmap to support more specific design and implementation decisions as states work toward their clearly specified end goals. These decisions lead to the development and implementation of the program, including (but not limited to): assessments, student accommodations, technical manuals, assessment resources (e.g., administration manuals, school reports and student reports), professional development, communication plans, and methods for gathering and using feedback for continuous improvement.

[4] Other examples of problems and goals being addressed by states via innovative assessment programs can be found here: https://knowledgeworks.org/resources/emerging-trends-k12-assessment-innovation/

Integrate Related Federal and/or State Program Goals into the Evaluation.
The evaluation design should support federal and state policies and programs related to the innovative assessment program. For example, states that participate in the IADA must comply with IADA regulations. The U.S. Department of Education specifies several IADA program goals that can be supported through a well-designed evaluation, including:

- Innovation in large-scale testing
- Evidence of stakeholder feedback
- Evidence of innovative program effectiveness (including sufficient educator training)
- A well-executed continuous improvement process
- Federal accountability (annual IADA program reporting)

Additionally, the federal IADA program requires annual reporting for federal progress-monitoring purposes, so federal officials can determine the extent to which states are making progress toward these goals. Thus, an evaluation of an innovative program under IADA would need to use both formative and summative evaluation methods to address goals (1) and (2) above. Moreover, the evaluation should examine the extent to which the program supports assessment innovation (goal 3) and is used for federal accountability purposes (goal 4). Non-IADA states may have their own set of program goals to be included in the evaluation design.

EVALUATION DESIGN AND IMPLEMENTATION PHASES

A useful evaluation provides feedback to inform program improvements and establishes evidence of a program's effectiveness. In the context of innovation, the evaluation should build evidence of the program's effectiveness in a planful way, from initial design through ultimate implementation at scale. Often, an innovative program may work well in one or a few sites, but the program fails to scale as intended soon after it expands to new sites or contexts.

A useful evaluation provides feedback to inform program improvements and establishes evidence of a program's effectiveness.

A program designed for use across a large population of sites or individuals—especially one as complex as a state assessment program—must be adapted over time as it diffuses into new sites (schools). The best and most useful evaluations account for this type of dynamic evolution by addressing two primary goals, which should target specific end-users (e.g., students, educators, schools):

A program designed for use across a large population of sites or individuals—especially one as complex as a state assessment program—must be adapted over time as it diffuses into new sites (schools).

- Informing the program's ***continuous improvement*** within a defined set of developmental phases; and
- Building ***evidence of the program's effectiveness*** over time, starting with a small and well-defined group of users (phase 1), continuing as the program is diffused across a broader group of users in less and less controlled environments (phase 2

and 3), and ultimately culminating in evidence that examines the program's effectiveness at scale (phase 4).

ESSA defines four tiers of evidence for establishing an intervention's effectiveness at scale. These tiers, designed for typical interventions, provide a helpful guide for building evidence of program effectiveness over time. Here, we adapt them to work with assessment programs. By designing an evaluation that aligns with these tiers of evidence, a state can make evidence-based claims about the effectiveness of its assessment program and, more specifically, its ability to achieve stated outcomes. Each of these tiers has implications for the development and implementation of an assessment program. Within the following subsections, we tie these tiers to phases of assessment program development and implementation.

ESSA defines four tiers of evidence for establishing an intervention's effectiveness at scale. These tiers, designed for typical interventions, provide a helpful guide for building evidence of program effectiveness over time.

ESSA's four tiers of evidence are defined below, followed by a brief description of how the evaluation supports progressive evidence within and across tiers.

- **Tier 4 (Demonstrates a Rationale).** The assessment program is supported by a well-defined theory of action that is informed by research. Additionally, the theory is evaluated by an outside research organization to determine its coherence with existing research and theories of learning. Any innovation must always begin at tier 4. That is, developers must start by creating a well-defined theory of action supported by research. Creating a theory of action is a critical component in the design process and one key reason why evaluating the design process itself should be included in any evaluation.
- **Tier 3 (Promising).** The assessment program is supported by one or more well-designed and well-implemented correlational studies. Tier 3 evidence is established by evaluating the innovation at a very small scale. Once the assessment program is ready—including the assessments and related training, materials, timelines, and implementation plans—the agency can test the approach with a small group of schools. Using internal and external evaluators, information can be collected and used to examine the implementation process and test relationships between implementation and associated outcomes of interest.
- **Tier 2 (Moderate).** The assessment program is supported by one or more well-designed and well-implemented quasi-experimental studies. Once the state agency (i.e., the state department of education) has sufficient evidence that the assessment program is working as intended, it can begin to scale the program to more schools. Additionally, information from both formative and more rigorous summative evaluations should be used to evaluate implementation and outcomes and inform continuous improvement to products and implementation processes.
- **Tier 1 (Strong).** The assessment program is supported by one or more well-designed and well-implemented randomized control trials (RCTs). Tier 1 occurs once the program is implemented at scale. In the context of an innovative assessment program, a tier 1 evaluation would likely not occur until at least several years after the pilot begins and potentially not until five years or more into the implementation.

Building an evidentiary base for these tiers requires both formative and summative evaluation. This means that throughout the development and implementation process, both formative and summative evaluation activities will need to be strategically and flexibly implemented. Formative and summative evaluations complement one another and provide useful information to inform improvement and establish evidence of efficacy over time. In the sections below, we describe the formative and summative evaluation process. We begin with a brief description of each (formative and summative) and follow with a list of key questions, methods and activities, and program outputs.

Develop the Formative Evaluation Approach

A ***formative evaluation*** is designed to produce feedback through iterative improvement cycles and prototyping. Formative evaluation enables assessment program designers to develop and continuously improve assessment program components. Information gleaned from a formative evaluation is used to improve the design, implementation, and/or intended use of the program of assessment program components.

Formative evaluation enables assessment program designers to develop and continuously improve assessment program components.

Questions. Questions that inform the formative evaluation tend to be more fine-grained than those in a summative evaluation; they focus on specific program components and inform the collection and analysis of information in frequent intervals. Additionally, questions are designed to quickly improve specific aspects, or components, of a program. Examples of questions in a formative evaluation might include:

- How are teachers using assessment reports?
- Are teachers interpreting score reports as intended?
- Are there aspects of the report that are difficult for teachers to understand?
- How and to what extent are teachers using assessment reports to inform their instruction?

For example, the North Carolina Department of Instruction (NCDPI) embedded a formative evaluation into its IADA program. During the second year of their innovative pilot program, NCDPI administered a teacher survey, conducted teacher focus groups, and solicited feedback after formal and informal presentations to district and school leaders from pilot schools. NCDPI leaders met regularly to review the qualitative feedback (i.e., focus groups and informal feedback collected via presentations/discussions with local leaders).

Findings were documented and iteratively updated as new information emerged. Surveys administered in the spring semester were used to both (1) verify the formative findings produced throughout the year and (2) identify areas of surprise (positive and negative) that required future investigation. This process enabled NCDPI to develop, test, and improve how information was presented in "NC Check-Ins 2.0" interim assessment reports for teachers' instructional purposes.

Additionally, a regular systematic data collection and review process allowed NCDPI to identify and address emerging challenges before they impacted statewide perceptions of the new personalized assessment tool being developed under IADA. On one occasion, the evaluator found that the state's communication plan did not include a strategy for communicating to non-pilot schools statewide about the new assessment program. As a result, NCDPI integrated presentations into conferences and other events during the following year to address the gap. Post-survey data is currently planned

so that NCDPI can monitor the extent to which these communication strategies are indeed promoting awareness and preparation for the new assessment program among all public schools.

Methods and Activities. A formative evaluation relies on iteration. Findings that emerge through frequent analysis cycles inform iterative stages of development (Bryk, 2015). These cycles are commonly referred to as "plan-do-study-act" cycles (PDSA), and they offer a helpful way to test and refine pieces of the overall program such as test blueprints, specific assessment item types, administration processes, score report designs, and professional development strategies.

A formative evaluation relies on iteration. Findings that emerge through frequent analysis cycles inform iterative stages of development (Bryk, 2015).

- **Plan.** The strategy being tested should be designed to address a root problem or key assumption/mechanism from the state's well-defined theory of action and logic model.
- **Do.** Implementing the plan with a small group of students, teachers, or in a small number of sites allows the state or district to determine whether a strategy is working as intended before expanding to new sites and, eventually, rolling out the program statewide. Other crucial activities include:
  - Systematically documenting how implementation was carried out;
  - Involving contextually diverse sites to understand where, and with whom, program outputs and outcomes vary; and
  - Collecting data on the efficacy of core program components
- **Study.** States and districts should collect data during implementation to systematically examine the changes that occurred, the extent to which those changes align with the underlying theory of action, and how changes vary across student subgroups and school sites. Additionally, data can be used to examine relationships between activities, outputs, and outcomes.
- **Act.** Using findings from the study phase, the study team can make decisions such as:
  - ***Adopting*** promising program components by expanding to new sites and continuing to test its efficacy.
  - ***Adapting*** components and strategies, which presupposes a change in the underlying theory of action or logic model. Adaptation may involve updating implementation procedures, continuing to test over longer periods of time, or modifying contextual factors that influence how program components unfold in a specific context.
  - ***Abandoning*** one or more program components or strategies and revisiting the problem statement and root causes to determine new solutions.

Once the cycle is complete, the original plan is revised or a new plan is created, and the cycle begins again. It is typically best to run at least three PDSA cycles before deciding to adopt, adapt, or abandon the specific practices, or program components, being tested. It takes time to fully understand why and how a strategy works when it is tested with a variety of individuals and under a variety of conditions (Shakman et al, 2020).

Organizations can implement different types of PDSA cycles (National Implementation Research Network, 2021):

- **Rapid-cycle improvement cycles** focus on resources or narrow aspects of a program. Rapid cycles occur quickly and on a small scale, often ranging from a few days to a few weeks. An example would be testing and then modifying a score report using cognitive think-alouds with teachers. Suggested changes to a score report can often be quickly incorporated into new reports, re-tested, and refined.
- **Usability testing cycles** involve more complex program components and therefore take longer than rapid cycles. Usability cycles range from a month to several months. The intent of these cycles is to understand how program components may vary based on student background or school context. For example, a state might conduct usability testing cycles to evaluate and improve assessment literacy training for elementary and middle school teachers. Because training consists of several resources and activities that are implemented over weeks or months, cycles of improvement are naturally longer.
- **Policy-practice communication cycles** tend to occur on a semi-annual or annual basis. These cycles require changes across multiple organizations or departments within a system. An example would be when a state examines the extent to which its strategies for communicating about a new assessment program reach its intended audiences.

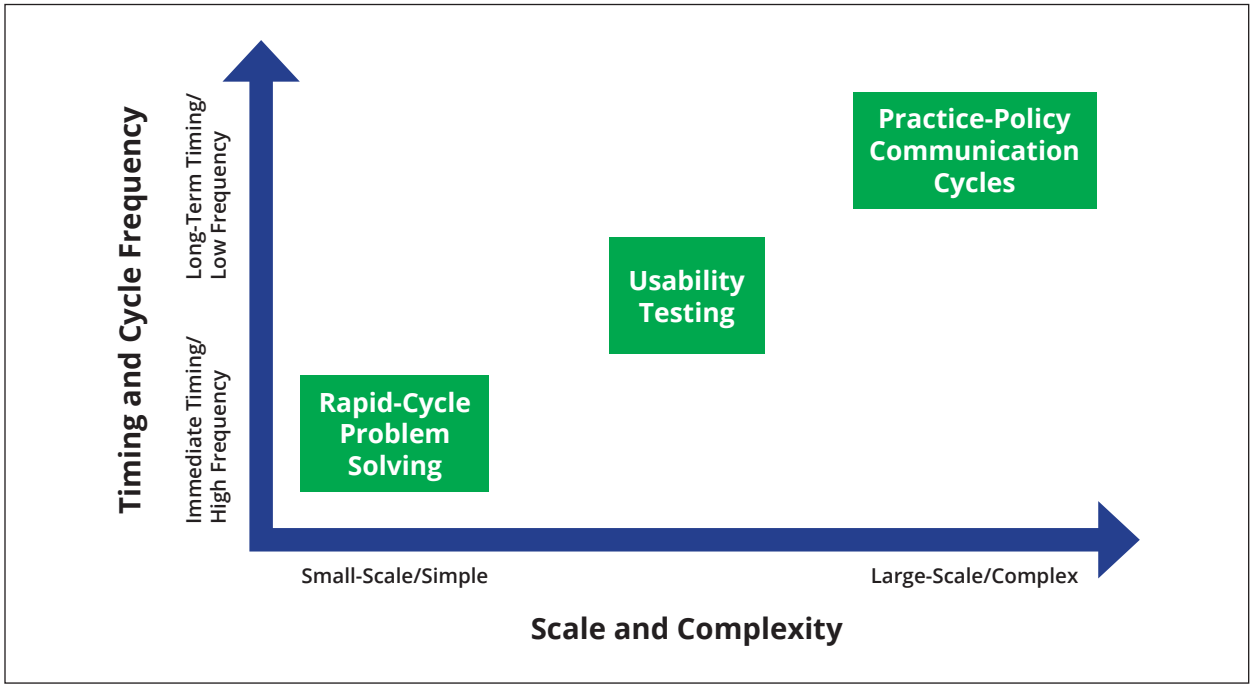
Notably, as improvement cycles increase, their purposes and uses may resemble features of a summative evaluation. Taking the example above, a state might conclude that its communication strategies for promoting the new assessment program worked well for educators but not for parents. This is a summative judgment, but nonetheless useful for guiding future improvements to improving communication the following year.

Figure 2 illustrates how the scale and complexity of an organization's focus will affect the type of PDSA cycle required. Typically, as the scale and complexity of an improvement effort increases, the time required for engaging in a full PDSA cycle increases. PDSA cycles focused on improving policies, updating resources, or improving communication practices within a state agency occur much more slowly and less frequently than PDSA cycles implemented in schools or classrooms. Complex organizational changes tend to occur more slowly than classroom or school-based changes. These PDSA cycles, as well as any formative evaluation strategy, can and should be applied at each phase of implementation in Figure 1, although the focus, scale and specific evaluation questions will change to reflect each phase's emphasis.

Leadership plays a key role in supporting both state- and district-level PDSA cycles, as well as school-level rapid PDSA cycles. Through thoughtful planning, states and districts can initiate evidence-based changes that are informed by these cycles of inquiry.

Different types of PDSA cycles may be implemented at the same time at different levels of the system. Leadership plays a key role in supporting both state- and district-level PDSA cycles, as well as school-level rapid PDSA cycles. Through thoughtful planning, states and districts can initiate evidence-based changes that are informed by these cycles of inquiry.

Figure 2: Common Types of PDSA Cycles by Scale/Complexity and Timing/Frequency



PDSA cycles can be coordinated to ensure coherence across levels of a system. We have seen PDSA cycles work well when adequate resources and structural supports, and clear processes are in place. Below we illustrate how one state department supports local PDSA cycles as part of a larger continuous improvement model.

The Virginia Department of Education's Office (VDOE) of School Quality adopted a networked improvement communities (NIC)-approach to support district and school improvement efforts across the state. School divisions volunteer to participate in the NIC. The VDOE "NIC Champion"—a school improvement coordinator at VDOE— oversees the NIC program. Through her oversight, VDOE staff partner with Region 5 Regional Comprehensive Center to collectively provide training for local district leaders. Ongoing monthly training educates leaders on the NIC process and provides clear guidance and support related to key steps in the process: (1) problem identification, (2) systems-mapping and root cause analysis, (3) theory of action and logic model development, (4) implementation and communication planning, and (5) evaluation.

The evaluation step involves data analysts at both the district and state level. Analysts work closely with district and school leaders to identify and map existing data sources – including leading and lagging indicators - to the program logic model. Data sources might include leading indicators such as teacher, student, and parent surveys; curriculum review results (using high-quality rubrics); school "walkthrough" checklists; attendance data; and discipline data.

Common lagging indicators include interim and summative assessment results, college ready assessments, and/or diagnostic tests. Analysts then serve a critical role in the NIC as data providers; they generate and produce reports for local, district, and state improvement monitoring over time.

At a local level, school teachers and leaders align their PDSA processes to address program objectives to complement district- and state-reported measures. Teams within the school collectively introduce a small change (e.g., introducing formative assessment practices within daily math lessons) and use "practical" measures – questions or observations that can be quickly and systematically gathered via Google Surveys or spreadsheets.

Practical measures are not always quantitative; they may include fieldnotes and narrative observations. The key is that teachers and schools are able to quickly collect, and commonly compare, information during or immediately after a change is introduced. Bi-monthly team meetings are used to review information and make decisions about whether to adopt, adapt, or abandon the change initiative.

Successful change initiatives are shared across the school and implemented on a broader scale, and the cycle continues. The lagging data from state and district sources provides helpful feedback for monitoring the change initiative's efficacy on a broader scale (e.g., at a school level, or from fall to spring) as the change is improved and scaled.

Key Outputs. The key outputs will vary depending on the specific questions asked in the PDSA cycle. For example, outputs from the questions described above might result in a revised assessment score report. This revised report might then be introduced in a new set of schools, and the process would begin again.

Suggested Program Deliverables for Formative Evaluation
A well-designed and executed formative evaluation process should result in a core set of deliverables. These deliverables often include an improved set of program resources, strategies, and practices, along with detailed documentation of the process that an evaluation team used to systematically gather information and used it to produce these improved resources.

Develop the Summative Evaluation Approach
A ***summative evaluation*** is designed to produce evidence of effectiveness of a program's effectiveness over time and can be established according to ESSA's four tiers of evidence. Each of these tiers corresponds to a phase of the implementation and evaluation of an assessment program. Each phase of an evaluation has a different set of goals and activities. The evaluation evolves as the assessment program develops and matures and is implemented in more and more sites. Key goals, questions, and activities of the evaluation are listed underneath each phase of the evaluation.

A summative evaluation is designed to produce evidence of a program's effectiveness over time and can be established according to ESSA's four tiers of evidence.

In Figure 1, the arrow connecting the formative to summative designs reflects two ideas: (1) the formative evaluation process is integrated into the overall evaluation at every phase of program development; and (2) the formative evaluation process should support and inform the summative evaluation. Regarding the second point, information about the measures, performance indicators, data collection protocols, and the program itself (i.e., effectiveness of individual program components) should inform decisions about when and how to expand both the program and the summative evaluation when the program is ready for the next phase of development.

Below, we describe key questions, methods, activities, and outputs within each of four phases of the evaluation, beginning with program design and continuing through evaluation at scale (statewide). Each of these phases is detailed with evaluation in mind, but also corresponds to an idealized progression of assessment program development, starting from the initial conceptualization of a program all the way to full scale implementation.

Program Design and Prototyping

The evaluation in phase one should be designed to accomplish two goals: (1) establish evidence of a program's theoretical rationale and (2) support the iterative development of the most essential program components for implementation.

Questions. Evaluation questions addressed in this phase may include:

- How were the program's theory of action, implementation plan, and program components (i.e., assessments and supporting materials) developed and designed? To what extent is the program's design coherent with existing research and modern theories of learning?
- To what extent do program developers facilitate an inclusive process and integrate needs and priorities of interested parties (e.g., subgroups) into the design?
- Are the program's theory of action, components (i.e., assessment designs, supporting materials) and implementation plan grounded in research-based evidence?
- Does the program meet the constraints of federal law (e.g., technical quality; summative determination of student proficiency; use of results within state's school accountability system; etc.)?
- Does the plan articulate a well-defined and defensible process for achieving stated goals?
- What evidence exists to support the technical quality of assessment products and materials associated with the program based upon intended use?
- Judging against industry standards, to what extent are assessment products and associated materials ready to pilot?

Methods and Activities. To establish evidence of a theoretical rationale, evaluators thoroughly document the state's process for identifying an assessment problem, developing a theory of action, and creating an implementation plan. Methods ideal for documenting evidence of "what happened and how it unfolded" include ethnographies, case studies, participant observation, interviews, surveys, and document review (e.g., meeting agendas and summaries). Multiple sources of information collected by these methods can then be used to address the evaluation's key questions.

To establish evidence of a theoretical rationale, evaluators thoroughly document the state's process for identifying an assessment problem, developing a theory of action, and creating an implementation plan.

Evaluation activities in this phase are primarily descriptive in nature. Evaluators' work focuses on documenting the development of the assessment program as it unfolds, such as how assessment tools and supporting materials/resources (e.g., administration manuals) are developed against industry standards of quality, and how evidence of a scientifically defensible plan to ensure program fidelity (e.g., professional development) is established. Additionally, evaluators develop, test, and refine a measurement infrastructure and data collection protocols to prepare for later phases of the

evaluation (i.e., pilot and efficacy studies). In this way, the formative evaluation informs both the program components and evaluation protocols and plans.

Program Outputs: The essential components of innovative assessment programs can generally be grouped into four strands:

- A set of assessments,
- Core materials required to properly administer, interpret, and use assessments for a pre-determined set of purposes,
- Professional development for end users to implement the program with fidelity
- Plans and materials for communicating the goals, purposes, and intended uses of the program and its core components (e.g., assessment tools, reports)

A primary purpose of the evaluation in the *program design and prototyping* phase is to make value judgments about program components. For example, are the program components—including the system of assessments, assessment materials, such as administration manuals and professional development materials and processes—ready for pilot?

To make these determinations, evaluations may refer to industry quality standards. For example, new performance tasks developed for the New Hampshire PACE Project went through a systematic principled assessment-development process, cognitive labs, and a careful multi-layered review process to establish initial evidence of quality before they were piloted with students (Center for Assessment, 2020). Similarly, to meet quality standards, a set of modular assessments must adhere to a specific development process before they are ready for pilot, including alignment to the depth and breadth of the state's challenging academic content standards, fairness and accessibility evaluations.

The Standards for Educational and Psychological Assessment (2014) serve as the source for evaluating the extent to which assessments meet standards of quality. Additionally, depending upon the innovative assessment program design, states may need to specify the process for how information from multiple assessments will be scaled together into a summative determination of student proficiency, and the extent to which the innovative assessments will need to be secure. Although industry standards of quality do not exist for core assessment materials, professional learning, or communication, many of these same steps apply to ensuring that ancillary materials, such as cognitive labs, expert reviews, simulated pilots, are ready to pilot

The Standards for Educational and Psychological Assessment (2014) serve as the source for evaluating the extent to which assessments meet standards of quality.

Below is a suggested list of program components (also called program outputs) that derive from the design process. The formative evaluation will continue to support improvements to program outputs at this stage, and the summative evaluation should provide sufficient evidence to determine when they meet a standard that is sufficient for the next phase of development: piloting the program in a small set of schools.

Suggested Program Deliverables in Phase One

- Well-defined problem statement
- Theory of action
- Implementation plan (timeline, assignments, implementation process, communication plan)
- Detailed documentation of the design process
- High-quality assessments and supporting materials (with quality defined by experts in the field and industry standards)

Program Pilot

After program design and prototyping, the evaluation focuses on establishing evidence of the program's promise for achieving its goals in a targeted set of districts, grade levels and subject areas. Assessment designers should be ready to pilot the assessment program, or major program components, in a small and representative sample of schools. Schools participating in the pilot should be familiar with, and generally supportive of, the new assessment program design.

In the program pilot phase the evaluation should continue to focus on whether program designers establish sufficient evidence of technical quality for core assessments in the program.

Additionally, evidence of validity, reliability, and fairness should continue to be collected during this phase. Program developers must collect additional evidence of technical quality required for federal peer review during a larger pilot. Thus, in the program pilot phase the evaluation should continue to focus on whether program designers establish sufficient evidence of technical quality for core assessments in the program.

Questions. Evaluation questions addressed in phase two include:

- To what extent were assessments administered as intended in the pilot schools?
- To what extent did assessments meet technical quality standards?
- How well did materials such as administration manuals support consistent administration across school sites?
- What improvements are needed?
- To what extent do resources and training support the intended uses of the assessment results?
- What improvements and adjustments are needed to ensure the technical quality of assessment results, and high-quality implementation, in order to meet the expectations of federal peer review?

Methods and Activities. In this phase, evaluators should have established methods for documenting the design and development process. Additionally, the evaluation expands to examine:

- Evidence of the technical quality of program assessments and summative determinations of student proficiency,
- Quality, relevance, and usefulness of training and resources to support assessment administration, interpretation, and use, and
- Hypothesized relationships between inputs, outputs, and outcomes specified in the theory of action.

To examine technical quality, evaluators may review the documented assessment evidence against federal peer-review requirements. The review enables the evaluation team to identify potential gaps where more information may be necessary to establish technical quality across one or more assessments. Additionally, evaluators can examine the technical quality of measures used in the evaluation.

Typically, pilots should include enough schools to facilitate examination of the extent to which surveys, interview protocols, and other data collection tools provided useful information to inform the innovative program. Evaluators can use data from the pilot to revise these tools before a larger efficacy trial ensues. In an efficacy trial, the program is diffused to a much larger population of schools and conditions vary more widely than in a pilot.

The pilot is also ideal for examining the quality, relevance, and usefulness of training and resources. For example, the IADA annual performance report (APR) requires states to collect stakeholder feedback about the extent to which the development process was collaborative and included a range of stakeholder groups. The APR also asks for feedback from teachers, school leaders, and parents about their "satisfaction with the innovative assessment system" (2021 APR, p. 10).

The pilot is also ideal for examining the quality, relevance, and usefulness of training and resources.

Surveys are often an efficient way for program participants to gather feedback on the program and its core program components. Interviews, focus groups, and observations complement survey results. These types of qualitative methods can be helpful for examining contextual factors that influence educators' perceptions. Additionally, qualitative methods provide deeper insights into the factors that facilitate and inhibit implementation.

Although pilots typically include a small number of schools and conditions tend to be tightly controlled, the evaluation team can apply inferential models to examine relationships between inputs, outputs, and outcomes. For example, evaluators can determine whether training and guidance to support assessment administration resulted in highly standardized administration across sites.

Additionally, evaluators can examine the extent to which training on the interpretation and use of results influenced expected changes in classroom-based behaviors—if that was part of the program's theory of action. For example, "To what extent did teachers who attended training sessions and accessed available resources make changes in their instruction after reviewing assessment results?" The pilot offers evaluators the opportunity to address these types of questions and revise the theory of action and/or specific program components to improve expected outcomes.

After a small pilot, a state will typically have sufficient evidence to expand it.

Program Outputs: The purpose of the evaluation in the pilot phase is to determine whether the program works under ideal conditions (i.e., the system of assessments, core materials, and professional development/training) and shows promise for scaling up. The summative evaluation will address whether the program can be (1) implemented with fidelity and (2) produce expected outcomes when it is introduced in a demographically representative set of schools. Evaluators

should have sufficient evidence to determine whether the program works as intended in a small group of schools and under tightly controlled environments. Once this bar is reached, the program can be scaled to more schools to examine efficacy across a variety of conditions.

Suggested Program Deliverables in Phase Two

- Refined theory of action
- Memorandum of understanding (signed commitment) from state and local agency stakeholders to support the innovative assessment program
- Updated implementation plan to scale the program statewide (timeline, assignments, implementation process, communication plan)
- Detailed documentation of the design and development process (expanded from phase one)
- Evidence of technical quality to meet federal peer-review standards
- Evidence that training and resources are of high quality, relevant, and useful
- Evidence establishing strong relationships between key program inputs, outputs, and outcomes in a small and representative sample of schools (e.g., sufficient representation across ESSA and state-established subgroups)

Program Expansion

Program expansion often consists of an efficacy trial. An efficacy trial determines whether an intervention produces the expected result *under ideal conditions* (Gartlehner et al., 2006). In the context of a state assessment program, this means that participants (e.g., schools, leaders, teachers) buy into the program's promise of success and are motivated to implement the program as designed.

An efficacy trial determines whether an intervention produces the expected result *under ideal conditions* (Gartlehner et al., 2006). In the context of a state assessment program, this means that participants (e.g., schools, leaders, teachers) buy into the program's promise of success and are motivated to implement the program as designed.

Schools in an efficacy trial, ideally, should be fully informed about implementation expectations; they commit to purchasing the necessary resources and establishing the conditions for success. Collectively, the sample of school participants should reflect the broader statewide population of schools, particularly with regard to the factors presumed to affect implementation (e.g., geography, organization, teacher and student demographics). That said, efficacy trials often begin by implementing the program in specific school sites, grade levels and/or subject areas and then expanding to new sites, grades and subject areas as evidence of program effectiveness builds over time. Ultimately, the goal of an efficacy trial is to determine whether the program works, for whom it works, and under what conditions it works before it is rolled out at scale.

Later stages of the efficacy trial in phase three may expand the program to cover new grades/subject and contexts that may have been underrepresented in the pilot phase. Additionally, while some grades/subjects may be ready for an efficacy trial, other grades can begin the pilot phase. For example, the pilot may show the program works as expected in elementary mathematics, but not in Algebra I at the high school level. If this happens, the state may decide to revisit the theory of action and may

require a new pilot that includes significant revision to the original assessment plan in high school.

Questions. Evaluation questions addressed in phase three include:

- Were test resources (e.g., schoolwide training, web-based materials) delivered by the responsible organization(s) as planned? Were resources received and used by school leaders and classroom teachers as planned?
- Did teachers make the intended instructional changes in their classrooms?
- Did the program (i.e., administration, reporting, and use of results) affect the performance of students?
- How did implementation and outcomes vary across school contexts and student subgroups?
- What are the minimum dosage levels needed for core components of the program (e.g., training participation) to bring about positive changes in outcomes such as instruction and student performance?

Methods and Activities. As the program is expanded, evaluators should be ready to test the program's efficacy using rigorous experimental or quasi-experimental methods. Additionally, district and school sites should be motivated and eager to implement the new program. The goal is to determine whether the program can work in less controlled environments with *willing* participants, so getting schools to buy into the innovative assessment program and fully participate in the evaluation is important. This allows the evaluation to address the next big question: "Does the program work when implemented across a diverse range of schools?"

As the program is expanded, evaluators should be ready to test the program's efficacy using rigorous experimental or quasi-experimental methods. Additionally, district and school sites should be motivated and eager to implement the new program.

Though not a common practice in most school contexts, the state may choose to randomly assign a subset of schools who are willing to participate in a formal study to experimental and control groups. The study can then investigate implementation and compare outcomes among those that received minimum support (e.g., just administration support) versus those that were exposed to the full program package (access to a range of classroom assessments, comprehensive training/coaching, curricular materials, etc.). If experimental trials are not feasible, the state may decide to use time series data to examine how instruction and student outcomes change over time in treatment vs. demographically similar comparison schools.

The more sophisticated the program, the longer it will take to establish a program's efficacy. Programs that include substantial training and resources will require more time for teachers and students to adapt. For example, decades of school reform suggest that it can take anywhere from three to 10 years for a program to take hold and change a system (Comer, 1980; Desimone, 2002). A state should plan to pilot and scale the innovative assessment program over multiple years, in order to test and refine it in targeted grades/subjects

The more sophisticated the program, the longer it will take to establish a program's efficacy.

and to establish strong evidence of effectiveness over time. Because of this, evaluation stakeholders (e.g., evaluators, internal staff, and/or community members involved in the change initiative) should focus on the highest leverage aspects of the program; those that are likely to produce the biggest bang for the buck.

Program Outputs. The goal of an efficacy study is to establish the program's efficacy under ideal conditions. Once this happens, then the program should be ready for implementation at scale.

Suggested Program Deliverables in Phase Three

- Evidence of the program's efficacy using rigorous evaluation methods (e.g., experimental and quasi-experimental designs) within and/or across a targeted set of grades and/or subjects
- Evidence that participants included a large and demographically diverse set of schools motivated to implement the new program
- Evidence of efficacy in program delivery, school implementation, classroom enactment, and student outcomes

Program at Scale

The summative evaluation at program scale examines the program's effectiveness at scale. In this phase, specific subjects and/or grade levels will be ready for a full-scale rollout and evaluation at different time intervals. Thus, similar to phases two (pilot) and three (efficacy), a an effectiveness trial may only target one subject area at one or a few grade levels within a given year. However, the evaluation focuses on the program's effectiveness at scale.

Questions. Evaluation questions addressed in phase four include:

- Does the program work at scale?
- For whom and under what conditions does the program work?
- What factors may be mediating or moderating outcomes?

Methods and Activities. Evaluators at phase four may expand existing methods and continue pursuing questions developed in phase three, only now applied to all schools. In the context of a statewide assessment program, randomized experiments would no longer be feasible since all schools have now adopted the innovative assessment program. However, evaluators may still apply rigorous methods; for example, using longitudinal data to test the effectiveness of the program over time by examining trends in related measures that fall outside the program such as NAEP, TIMMS, or PISA.

They can also continue measuring fidelity of implementation, using findings in a formative way to improve program components or identify sites and/or subgroups that may need additional support. Well-designed effectiveness trials include analyses that examine factors that mediate performance outcomes such as schools' use of assessment reports, changes in classroom instruction, or changes in other school- and classroom-based practices. They also include analysis to examine how demographic and other static characteristics moderate performance outcomes.

Program Outputs. Effectiveness studies provide information to inform the variability of implementation and outcomes across sites and subgroups. This can happen through well-designed studies that investigate implementation fidelity and outcomes across representative samples of

schools and subgroups. The first key output would include information, reported via user-friendly reports, that communicates the effectiveness of the program statewide and for relevant subgroups. A second key output would include an iterative action plan that informs how resources and supports are used to improve the program's reliable impact across all schools and subgroups.

Suggested Program Deliverables in Phase Four

- Evidence of the program's effectiveness via annual reporting within and/or across relevant grades and/or subjects
- Reporting tools that facilitate the use of study results to target program improvements and support program efficacy across districts and school sites
- Action plans that communicate how resources will be deployed and support will be provided to improve program efficacy at scale, particularly for lower-performing schools and subgroups

Effectiveness studies provide information to inform the variability of implementation and outcomes across sites and subgroups. This can happen through well-designed studies that investigate implementation fidelity and outcomes across representative samples of schools and subgroups.

CONCLUSION

This paper presented a framework state education agencies can use to design and implement program evaluations of state assessment programs. Although our framework focuses on evaluating innovative assessment programs, the steps we outline can be used to guide evaluations of any program. The framework provides a roadmap that can help education agencies use evaluation for continuous improvement. It does this by relying on principles of improvement science (plan-do-study-act cycles) to illustrate how education agencies can improve their assessment programs. Additionally, it describes how agencies can establish summative evidence of program efficacy and effectiveness from inception through full-scale implementation.

Evaluating innovative assessment programs is particularly important right now. State assessment programs are more complex than ever, incorporating score reports for multiple audiences, online testing platforms, coupled interim assessments, corresponding assessment literacy efforts, and multi-faceted communication strategies. This complexity makes it all the more important to build a body of evidence, based on sound evaluation principles, that provides valid, reliable, and timely information about what is and is not working optimally. Moreover, federal incentives such as IADA and multi-year assessment grants have placed assessment at the center of state and local improvement efforts.

More and more education agencies are testing new and promising assessment innovations. As this continues, strong evaluations hold the promise of providing essential information to spread effective innovations, establish what works, and improve innovative assessment programs over time.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Bryk, A.S., Gomez, L.M., Grunow, A., and LeMahieu, P.G. (2015). *Learning to improve: How America's schools can get better at getting better.* Harvard Education Press.

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

Center for Assessment. (2020). *Performance Assessment of Competency Education (PACE): Evaluating technical quality manual (volume 1).* National Center for the Improvement of Educational Assessment. https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/pacetechmanualvol1.pdf

Comer, J. P. (1980). *School power: Implications of an intervention project.* Free Press.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research,* 72(3), 433-479. https://doi.org/10.3102/00346543072003433

Education First (2023). *What are through-year assessments? Exploring multiple approaches to through-year design.* Education First. https://www.education-first.com/wp-content/uploads/2023/01/What-are-Through-year-Assessments-1.pdf

Elementary and Secondary Education Act, § 200.104 (2015). https://www.ecfr.gov/current/title-34/subtitle-B/chapter-II/part-200/subpart-E/subject-group-ECFR9277b2b0db822d9/section-200.104

LeMahieu, P.G., Bryk, A.S., Grunow, A. and Gomez, L.M. (2017), Working to improve: Seven approaches to improvement science in education. *Quality Assurance in Education,* Vol. 25 No. 1, pp. 2-4. https://doi.org/10.1108/QAE-12-2016-0086

Gartlehner, G., Hansen, R.A, Nissman, D., Lohr, K.N., and Carey, T.S. (2006). *Criteria for distinguishing effectiveness from efficacy trials in systemic reviews. technical reviews, No. 12.* National Institutes for Health: Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK44029/

National Implementation Research Network. (2021). *Module 5: Improvement cycles.* National Implementation Research Network. https://implementation.fpg.unc.edu/wp-content/uploads/Improvement-Cycles-Overview.pdf

Shakman, K., Wogan, D., Rodriguez, S., Boyce, J., and Shaver, D. (2020). *Continuous improvement in education: A toolkit for schools and districts (REL 2021-014).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. http://ies.ed.gov/ncee/edlabs

King, J.B. (2016, February 2). *Follow-up on President Obama's testing action plan.* United States Department of Education. https://oese.ed.gov/files/2020/07/16-0002signedcsso222016ltr.pdf

United States Department of Education (2021). *2021 Annual performance report.* Washington, DC: United States Department of Education. https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/iada/

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2(2), 1-3. https://scholarworks.umass.edu/pare/vol2/iss1/2

APPENDIX A: AN OVERVIEW OF PROGRAM EVALUATION PHASES

Table 1: An Overview of Program Evaluation Phases

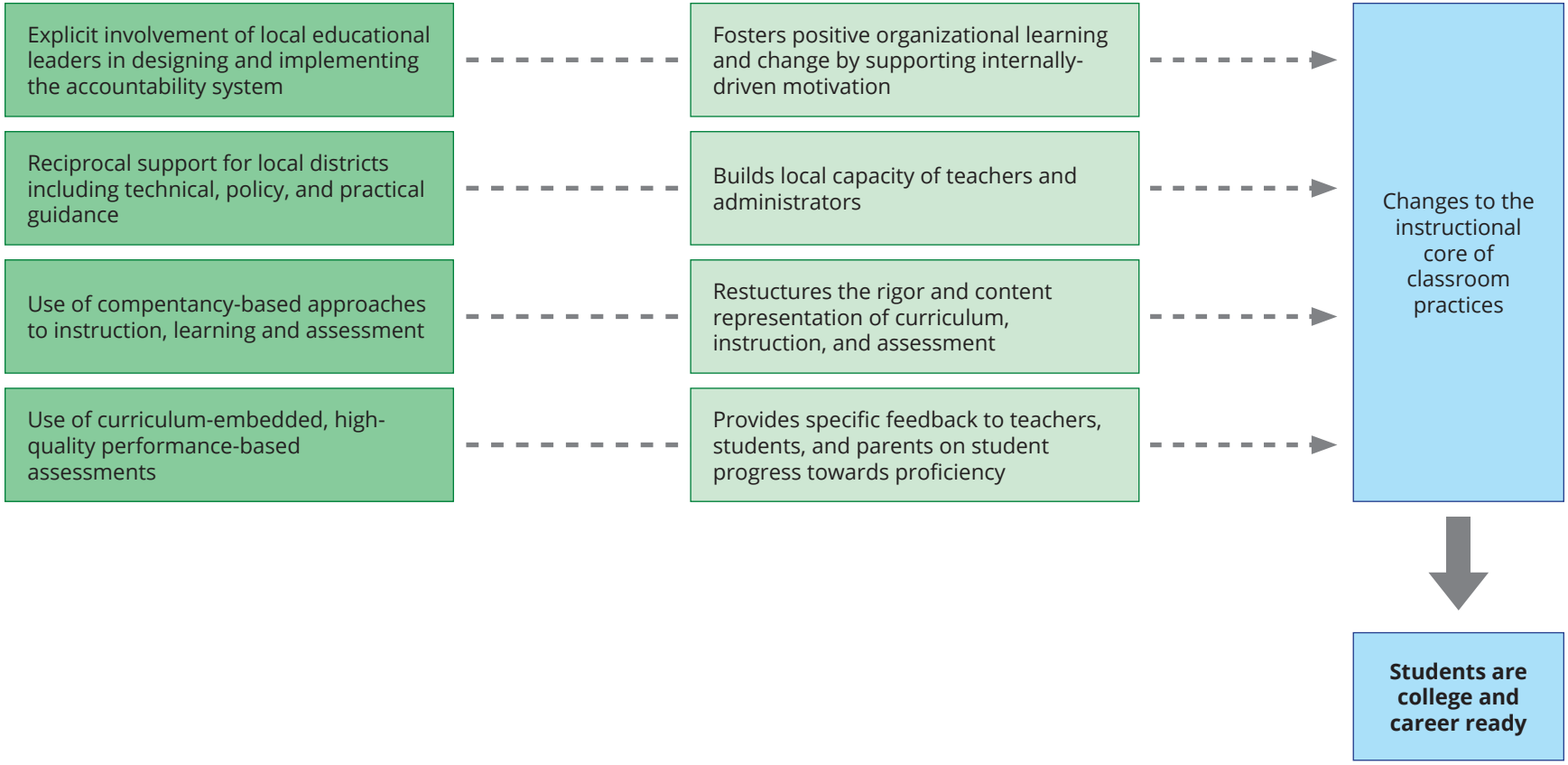
PHASE	GOAL	PRIMARY QUESTION	CORE ACTIVITIES	TIER OF EVIDENCE AND EVALUATION DESIGN PRIORITIES
Planning	Define the Program	• What is the program? • What is it meant to accomplish?	• Establish a clear vision • Explain what problem(s) the program is meant to solve and the goals and use cases connect to these problems • Articulate how the program is meant to function, typically captured through a theory of action	**Planning.** The program vision and theory of action should include a broad range of stakeholders representing a full range of expertise, geographies, demographics, and perspectives (e.g., policymakers, practitioners, community leaders, students). To the extent possible, the program vision and design should be grounded in research-based evidence of student learning.
Program Design and Prototyping	Demonstrate a Strong Theoretical Rationale	• Do developers establish a scientifically defensible theoretical rationale to address program goals?	• Describe the program • Document program goals and development process • Examine technical quality of program assessments • Examine stakeholder perceptions (e.g., Materials, PD) • Develop and refine evaluation measures and protocols.	**Tier 4 (Demonstrates a Rationale).** The assessment program is supported by a well-defined theory of action that is informed by research. Additionally, the theory is evaluated by an outside research organization to determine its coherence with existing research and theories of learning. Any innovation must always begin at tier 4. That is, developers must start by creating a well-defined theory of action supported by research. Creating a theory of action is a critical component in the design process and one key reason why evaluating the design process itself should be included in any evaluation.
Program Pilot	Establish Promising Evidence	• Do assessments meet technical quality standards (peer review)? • Does the program work in a small number of highly controlled contexts?	• Describe program activities • Document the development process • Establish technical quality of program assessments • Examine key relationships in the program theory of action • Establish initial evidence of the program's efficacy in a small number of controlled settings • Continue refining evaluation measures and protocols	**Tier 3 (Promising).** The assessment program is supported by one or more well-designed and well-implemented correlational studies. Tier 3 evidence is established by evaluating the innovation at a very small scale. Once the assessment program is ready—including the assessments and related training, materials, timelines, and implementation plans—the agency can test the approach with a small group of schools. Using internal and external evaluators, information can be collected and used to examine the implementation process and test relationships between implementation and associated outcomes of interest.

PHASE	GOAL	PRIMARY QUESTION	CORE ACTIVITIES	TIER OF EVIDENCE AND EVALUATION DESIGN PRIORITIES
Program Expansion	Establish Efficacy	• Does the program work in more diffuse contexts under ideal conditions	• Establish causal evidence of the program's efficacy in a larger number of ideal but less controlled settings • Examine how implementation fidelity diffuses in less controlled settings and specific contexts • Establish minimum dosage levels for the program and its major components • Continue refining evaluation measures/protocols for large-scale use	**Tier 2 (Moderate).** The assessment program is supported by one or more well-designed and well-implemented quasi-experimental studies. Once the state agency (i.e., the state department of education) has sufficient evidence that the assessment program is working as intended, it can begin to scale the program to more schools. Additionally, information from both formative and more rigorous summative evaluations should be used to evaluate implementation and outcomes and inform continuous improvement to products and implementation processes.
Program at Scale	Establish Effectiveness at Scale	• Does the program work at scale?	• Establish causal evidence of effectiveness at scale • Examine program fidelity and outcomes • Continue refining evaluation measures/protocols for large-scale use	**Tier 1 (Strong).** The assessment program is supported by one or more well-designed and well-implemented randomized control trials (RCTs). Tier 1 occurs once the program is implemented at scale. In the context of an innovative assessment program, a tier 1 evaluation would likely not occur until at least several years after the pilot begins and potentially not until five years or more into the implementation.

APPENDIX B: EXAMPLE THEORIES OF ACTION FOR INNOVATIVE STATE ASSESSMENT PROGRAM

New Hampshire Performance Assessment of Competency Education (PACE)[5]

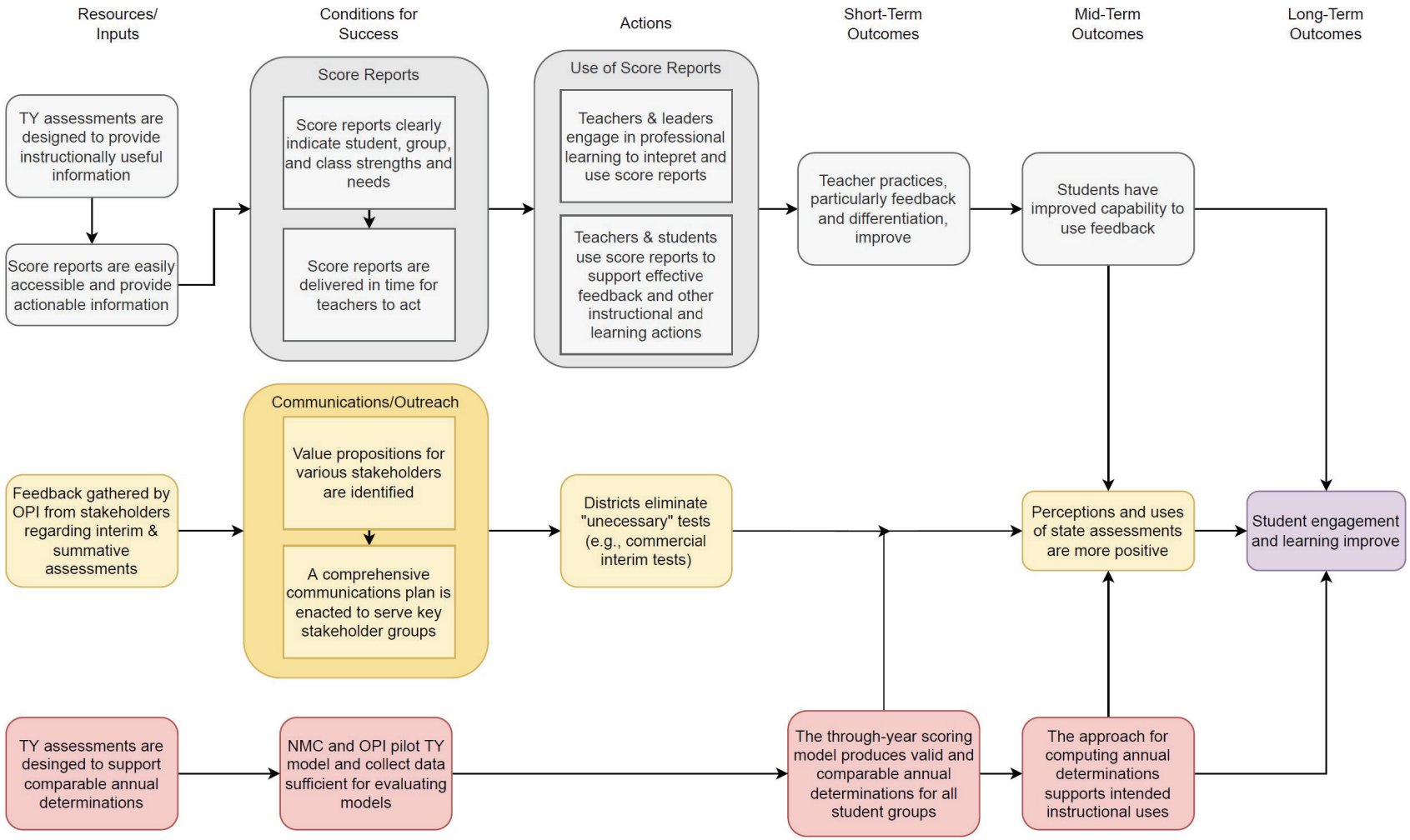
Figure 1. NH PACE Theory of Action



[5] For a comprehensive description, see: Lyons, S., Evans, C., Marion, S., and Thompson, J. (2017). New Hampshire Performance Assessment of Competency Education (PACE) Technical Manual. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://www2.ed.gov/policy/elsec/guid/stateletters/nhpacetechmanual72017.pdf

Montana Alternative Student Testing (MAST)[6]

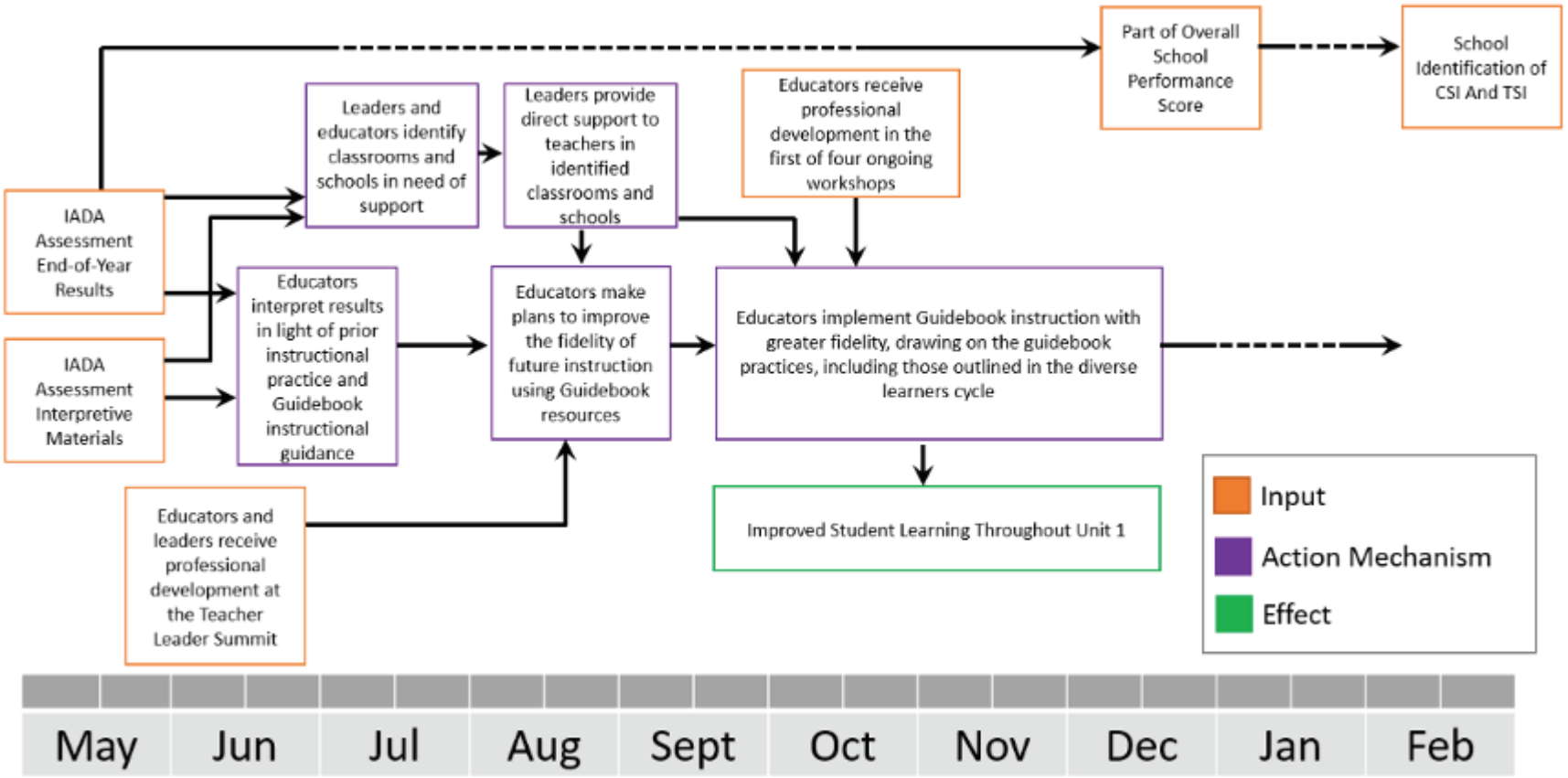
Figure 1. Overall Theory of Action



[6] For a comprehensive description, see: Marion, S., D'Brot, J., and Brandt, W.C. (2022). Assessment Design and Implementation Considerations for the Montana Alternate Student Testing (MAST) Pilot Program. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://sites.google.com/opiconnect.org/montanataskforceformathandelat/final-report

Louisiana Innovative Assessment Program (IAP)[7]

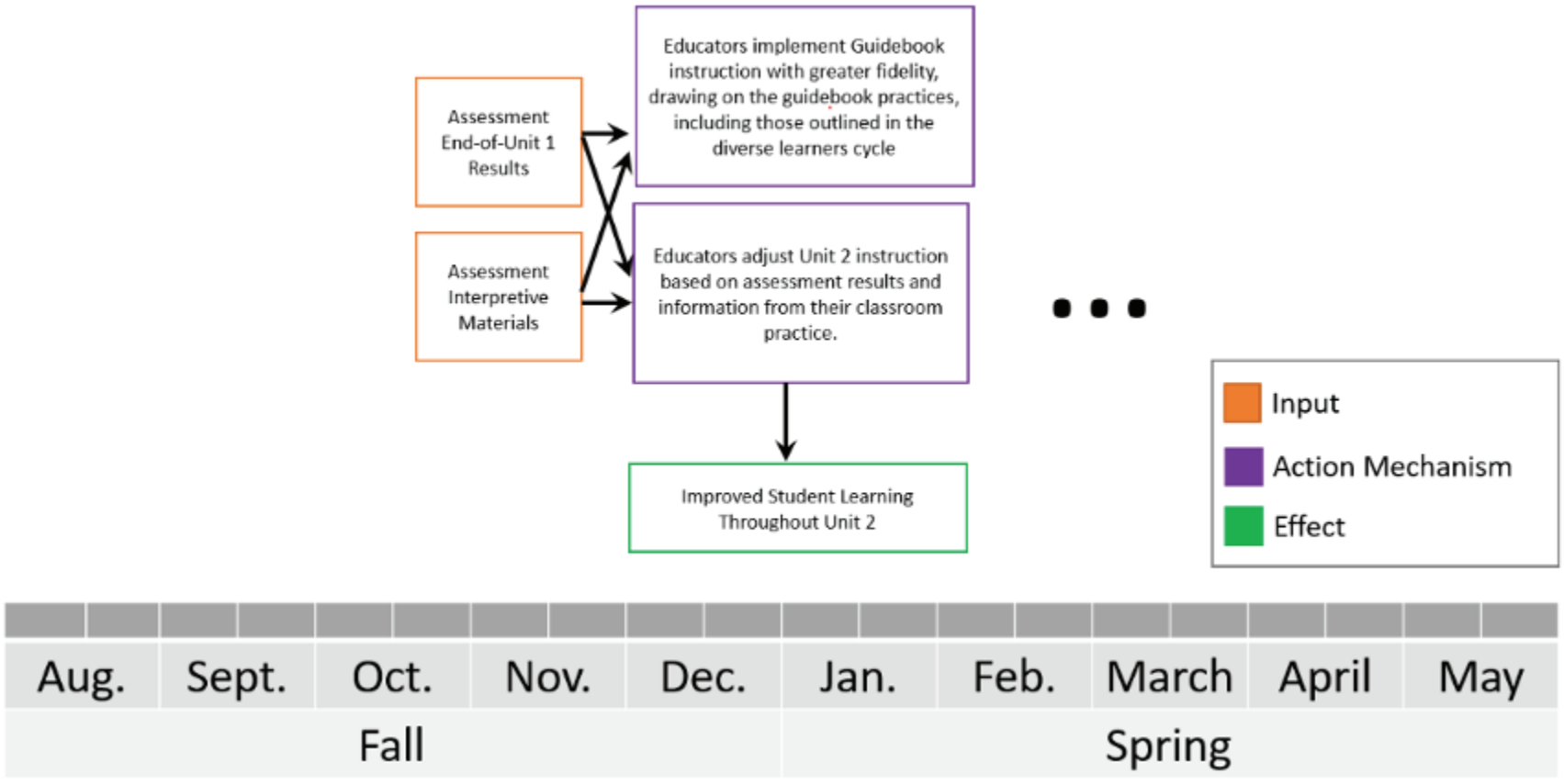
Figure 1. Working Summative Logic Model



[7] For a comprehensive description, see: Louisiana Department of Education (2021). Annual Innovative Assessment Demonstration Authority (IADA) Annual Performance Report. from the Annual Progress Report (p.127-128). Louisiana Department of Education. Retrieved from https://oese.ed.gov/files/2022/09/LADOE-IADA-APR-2020_21.pdf

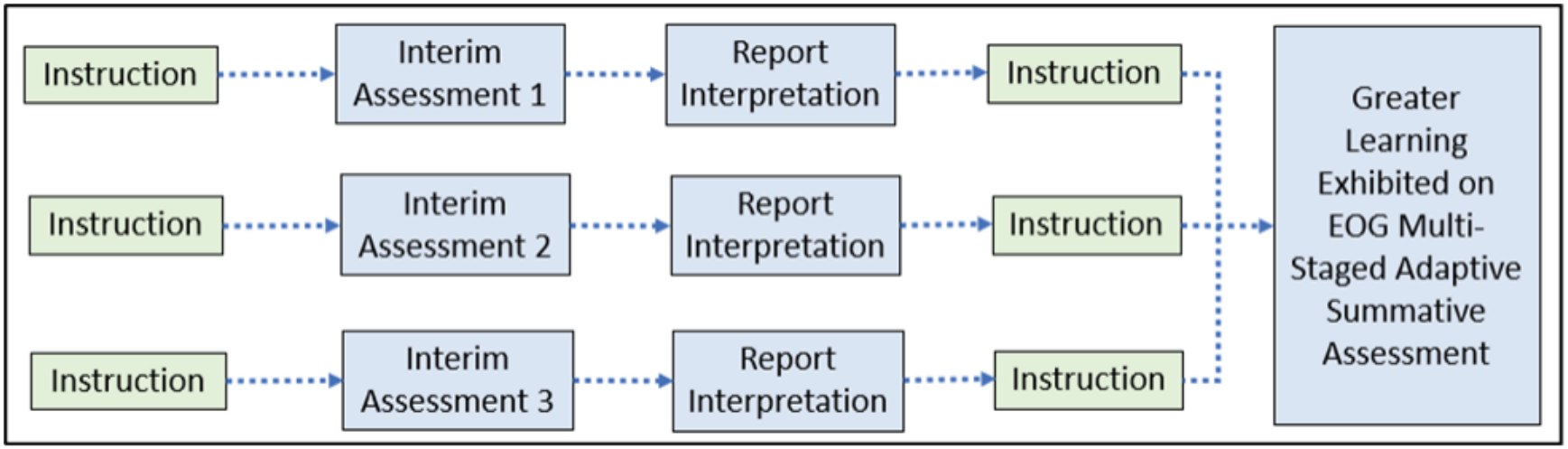
Louisiana Innovative Assessment Program (IAP)[7]

Figure 2. Working End-of-Unit Logic Model



North Carolina Personalized Assessment Tool (NCPAT)[8]

Figure 1. Program Overview



[8] For a comprehensive description, see Brandt, W.C. (2022). External Evaluation of North Carolina's Innovative Assessment Demonstration Authority (IADA) Pilot Program: The North Carolina Personalized Assessment Tool (NCPAT). Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from https://oese.ed.gov/files/2022/09/NCDPI-IADA-APR-21_22-1-1.pdf (p. 76-118).

North Carolina Personalized Assessment Tool (NCPAT)

Figure 2. Theory of Action

Goal *What is the overarching **goal(s)** of the system?*	**Outcomes** *What specific **outcomes** represent goal attainment?*	**Elements/ Components** *What **approaches**, **initiatives** and **components** need to be in place to support the attainment of outcomes?*	**Mechanisms** *What is the **mechanism** by which each element of the system will support the attainment of desired outcomes?*	**Assumptions** *What **assumptions** underlie the system working as intended?*	**Evidence** *What **evidence** will demonstrate that the system is working as intended?*	**Consequences** *What are the potential intended/unintended **consequences**?*
Intentional through-grade use of assessment data to support teaching and increase student achievement	A balanced assessment system consisting of formative, interim, and summative measures Increased achievement (short term/long term) Reduced achievement gaps Increased assessment and data literacy	Through-grade assessments (interims) Staged-adaptive summative Assessment of higher order thinking skills Professional development in assessment literacy with a common language of formative assessment Immediate teacher feedback Student reports	Variety of item types (e.g., TEI, performance tasks) Online reporting Professional development via training modules that can be accessed at any time: o Regional coaching o Online PD modules on assessment and data literacy o Online PD modules on the assessment system o Training on misconceptions	Data will be reviewed and used by educators. The system will provide valid and reliable data. The test is aligned to content standards. Teachers will integrate their increased understanding of assessment and data into their day-to-day practices.	Increased student achievement and growth o Higher percentage of districts meeting long-term goals (designed to close achievement gaps) (links to plans – ESSA, SBOE) o Reduction of low-performing schools, districts, and charter schools (link to SBOE)	Intended: Students have more timely feedback on their performance so that they can improve. Teachers have actionable information so that they can use it to change instruction for students. Unintended: Interims become high stakes. Increased stress around testing Testing perceived as increased testing (interims) Impact on local pacing guides

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

www.nciea.org