EVALUATING AND CONTINUOUSLY IMPROVING

an Innovative Assessment and Accountability System

October 2017

KnowledgeWorks

Center for Assessment

nellie mae EDUCATION FOUNDATION

Acknowledgements

Thanks to generous support from the Nellie Mae Education Foundation, KnowledgeWorks, and the National Center for the Improvement of Educational Assessment (Center for Assessment) have partnered to help states better understand and leverage the new Innovative Assessment and Accountability Demonstration Authority authorized under the Every Student Succeeds Act (ESSA). The goal of this partnership is to help states identify and explore a set of readiness conditions that are critical to the development of a high quality application and implementation process under this new authority. While we share a history of advocacy for next generation assessments, our organizations each bring a unique perspective to this work. KnowledgeWorks focuses on policy development, partnering with states, districts, and educators to identify and remove policy barriers that inhibit the growth of personalized learning. The Center for Assessment specializes in the design of assessment and accountability systems, helping states, districts, and other entities improve the quality of these systems and maximize student success.

Marion, S.F., Lyons, S., & Pace, L. (2017). Evaluating and continuously improving an innovative assessment and accountability system. www.innovativeassessments.org.

Table of Contents

Introduction

This is one in a series of policy and practice briefs produced by KnowledgeWorks and the National Center for the Improvement of Educational Assessment (Center for Assessment) designed to assist states in thinking through the opportunities and challenges associated with flexibility provided under the Every Student Succeeds Act (ESSA).[1] These briefs help define "Readiness Conditions" for states considering applying for and successfully implementing an innovative assessment and accountability system as defined by the Demonstration Authority opportunity under Section 1204 of ESSA.

Creating a State Vision to Support the Design and Implementation of an Innovative Assessment and Accountability System

Ensuring and Evaluating Assessment Quality for Innovative Assessment and Accountability Systems

Addressing Accountability Issues Including Comparability in the Design and Implementation of an Innovative Assessment and Accountability System

Supporting Educators and Students through Implementation of an Innovative Assessment and Accountability System

Evaluating and Continuously Improving an Innovative Assessment and Accountability System

Establishing a Timeline and Budget for Design and Implementation of an Innovative Assessment and Accountability System

Building Capacity and Stakeholder Support for Scaling an Innovative Assessment and Accountability System

[1]Brief five in a series of policy and practice briefs designed to help states prepare for the ESSA Assessment and Accountability Demonstration Authority. We are grateful to the Nellie Mae Foundation for their generous support of this project.

Overview

ESSA provides an opportunity for states to develop innovative assessment and accountability systems to evaluate student and school performance. Congress kept the language for this demonstration authority intentionally vague to ensure that states and districts have the space to innovate and design systems that align to their local vision and needs. Since there is not one "right" way to build an innovative assessment system, it is critical that states and districts establish high-quality processes to improve their systems along the way to maximize positive outcomes for students.

The final regulations for the Innovative Assessment Demonstration Authority, released by the U.S. Department of Education in December 2016, emphasize the importance of a high-quality continuous improvement and evaluation process. Specifically, the regulations include a set of five selection criteria that state applicants must address in their application for the pilot program. These selection criteria include:

1. **Project Narrative**
2. **Prior Experience, Capacity, and Stakeholder Support**
3. **Timeline and Budget**
4. **Supports for Educators, Students, and Parents**
5. **Evaluation and Continuous Improvement**

As such, states and districts will need to develop a robust plan and set of strategies to ensure stakeholders learn from, and continually improve their innovative assessment systems.

U.S. Department of Education Final Regulations

Every Student Succeeds Act—Innovative Assessment Demonstration Authority. (December 8, 2016)

Excerpt from § 200.106 Demonstration authority selection criteria.

(e) Evaluation and continuous improvement. The quality of the SEA's or consortium's plan to annually evaluate its implementation of innovative assessment demonstration authority. In determining the quality of the evaluation, the Secretary considers—

(1) The strength of the proposed evaluation of the innovative assessment system included in the application, including whether the evaluation will be conducted by an independent, experienced third party, and the likelihood that the evaluation will sufficiently determine the system's validity, reliability, and comparability to the statewide assessment system consistent with the requirements of §200.105(b)(4) and (9); and

(2) The SEA's or consortium's plan for continuous improvement of the innovative assessment system, including its process for-

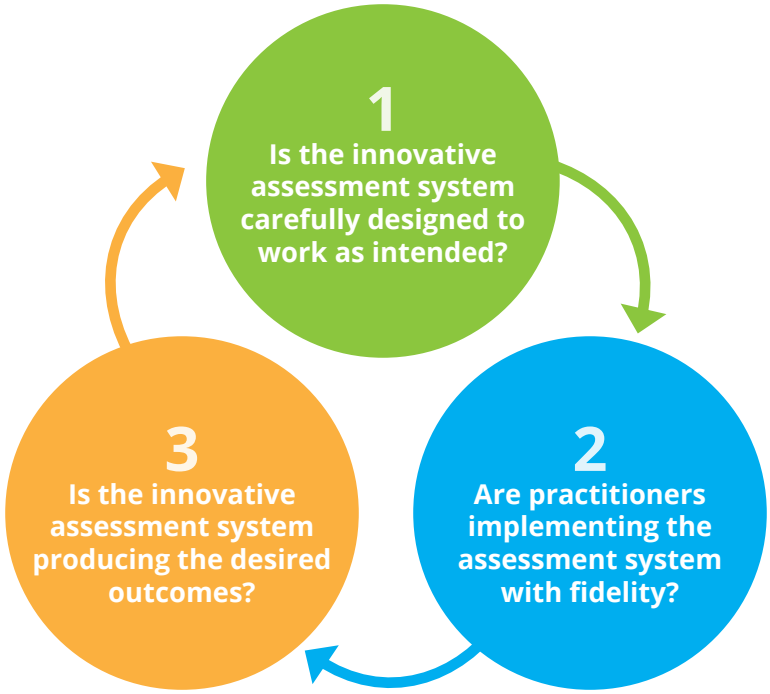
(i) Using data, feedback, evaluation results, and other information from participating LEAs and schools to make changes to improve the quality of the innovative assessment; and

(ii) Evaluating and monitoring implementation of the innovative assessment system in participating LEAs and schools annually.

State leaders and key stakeholders planning to apply for the innovative assessment pilot will do so because they expect it to result in improved student learning outcomes. But how would one know if this were the case? Judging the efficacy of any educational reform is a considerable challenge especially if one is hoping to see long-term or distal effects such as improved student learning. A key aspect of this challenge is that it requires evaluating at least three sets of assumptions. One set of assumptions is about the quality and efficacy of the design of the innovative system, a second set has to do with the fidelity of implementation of the designed system, and the final set addresses the observed outcomes of the system. All three sets of assumptions are rarely upheld; at least not fully. We must be humble about the design of truly innovative systems and we must recognize that implementing an innovative system will be a significant challenge for current practitioners. In other words, if the system is truly innovative, it means we do not have a lot of experience on which to base our design and implementation. Formative evaluation can provide "along the way" feedback on the innovative pilot to allow for early course corrections rather than waiting for the end of the pilot period to evaluate the effectiveness of the innovative system of assessments. This brief is designed to help educational leaders unpack key design and implementation assumptions and actions to work towards creating an innovative pilot informed by ongoing evaluation and continuous improvement activities.

Three Guiding Questions

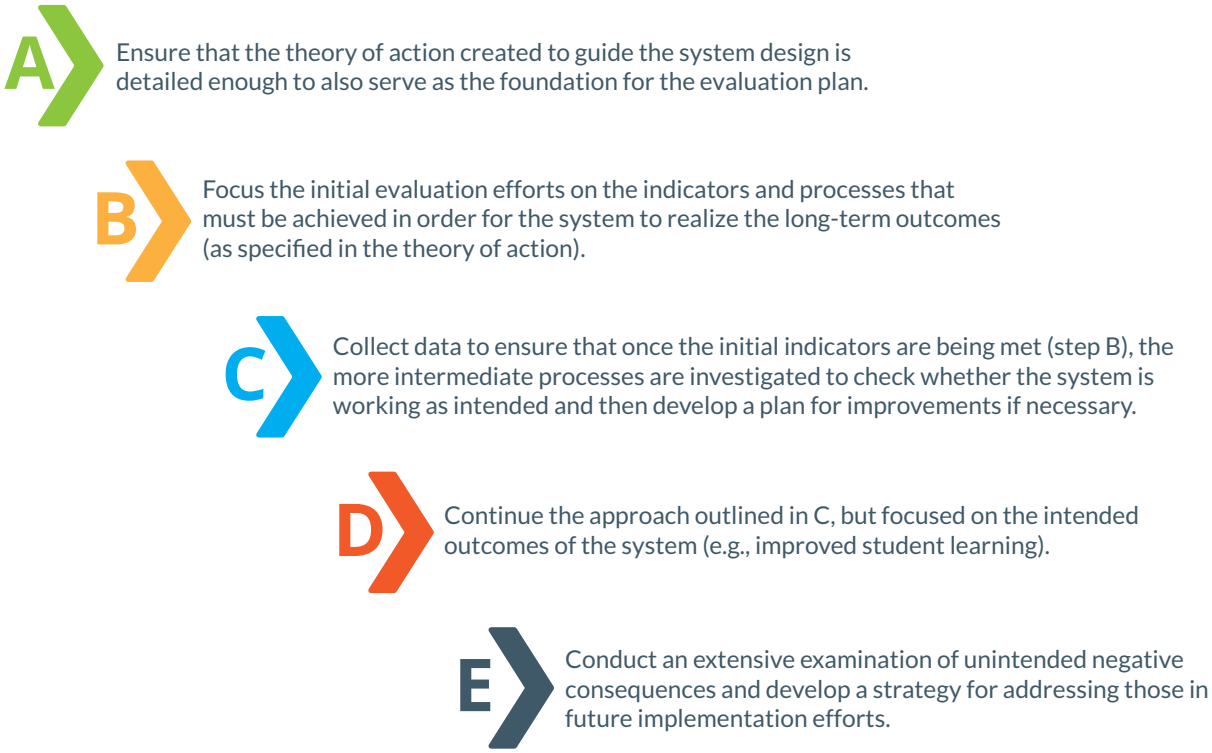
Building a continuous improvement system requires leaders to focus on three sets of questions:



Affirmative answers to questions 1 and 2 are necessary for producing the intended outcomes, but do not guarantee that the system will produce the desired results. Even when the design is based on a strong conceptualization, the innovation could be so novel that there is little experience on which to draw and the system breaks down in the implementation. As we have argued in earlier briefs, the innovation should be supported by a research base to the extent possible, but it could be the case that the research is not robust enough to accurately predict how the model works in practice. In such cases, state and district leaders need to rely on their evaluation of best practices and engage in a careful design process as discussed in earlier briefs. While it might be possible for the assessment system to produce the intended outcomes without careful examination of the first two questions, that outcome is more likely due to a case of good luck rather than disciplined, sustainable, and replicable innovation. Therefore, we argue that both the design and implementation of the innovative system are necessary conditions for meeting the intended goals of the system. As such, most of this brief focuses on the importance of on-going formative evaluation of the design and implementation of the innovative assessment system in lieu of a too narrow focus on outcomes alone.

Getting Started

It may appear overwhelming to both design an innovative assessment system and a formative evaluation process that will ensure ongoing feedback and improvement. We suggest that state leaders engage in the following steps as they design their evaluation and continuous improvement process.



We expand on each of these steps in the remainder of the paper. Importantly, step D and to some extent step E are designed to address the third guiding question, while the other steps will help address the first and second guiding questions.

Theory of Action

We previously published a brief about creating a theory of action to guide the design of the innovative pilot.[2] In that brief, we focused on the use of a theory of action for guiding the design of the system. Another key reason for creating a theory of action is to guide the formative evaluation of the designed system. That is our focus here. We revisit the examples of both the high-level and more detailed theories of action presented in our earlier brief to serve as touchpoints for our discussion that follows. Figure 1 is helpful for beginning to outline the system and to help align stakeholders behind the vision. However, for the theory of action to truly guide the design and evaluation of the innovative system, we must add considerably more detail about the processes and mechanism by which we expect the system to be realized.

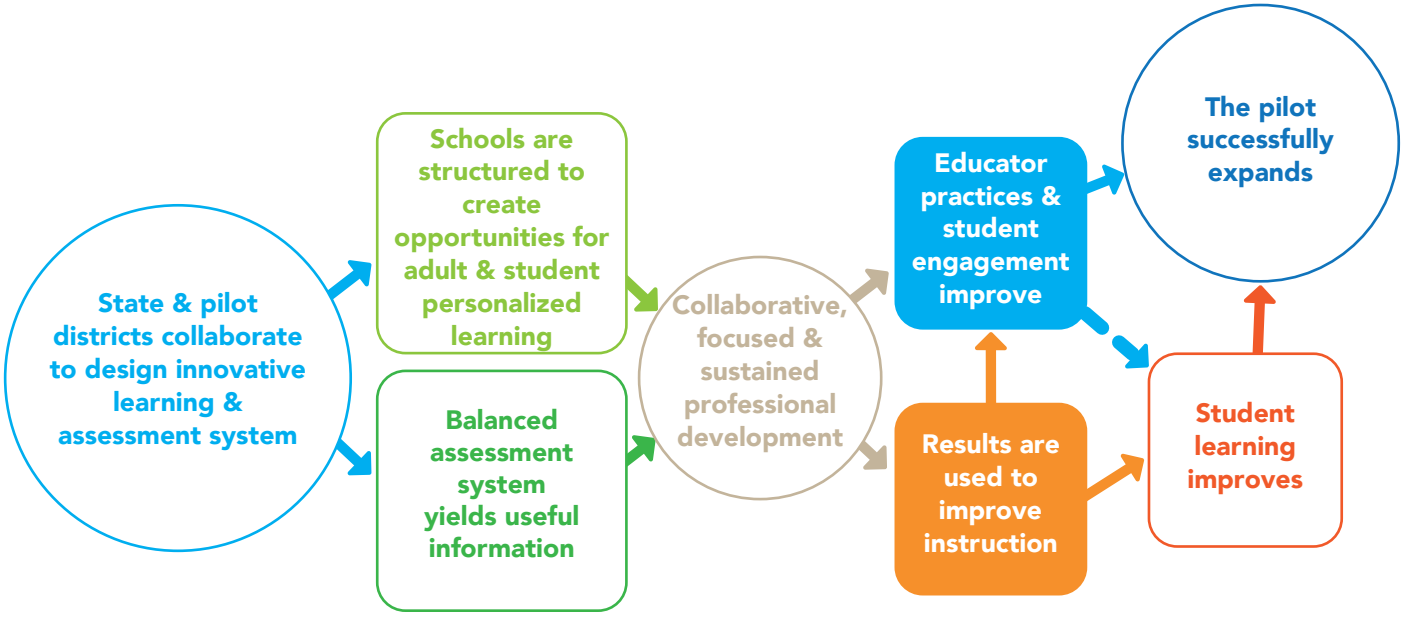


Figure 1. A theory of action for *improving practices and learning*

In Figure 2, we unpack just one aspect of Figure 1—the two shaded boxes—to illustrate the various considerations for state leaders and partners as they engage in this work. The shaded components of the theory of action from Figure 1 suggest that the assessment results are used to improve instruction, which then leads to improvements in educator practices and student engagement. Moving from assessment results to improved practices is no more than a leap of faith unless the designer articulates critical processes and activities to realize these intermediary outcomes. We provide an example of how state leaders and partners could begin unpacking these steps in Figure 2 below.

[2]Marion, S.F., Lyons, S., Pace, L., & Williams, M. (2016). A Theory of Action to Guide the Design and Evaluation of States Innovative Assessment and Accountability System Pilots. www.innovativeassessments.org.

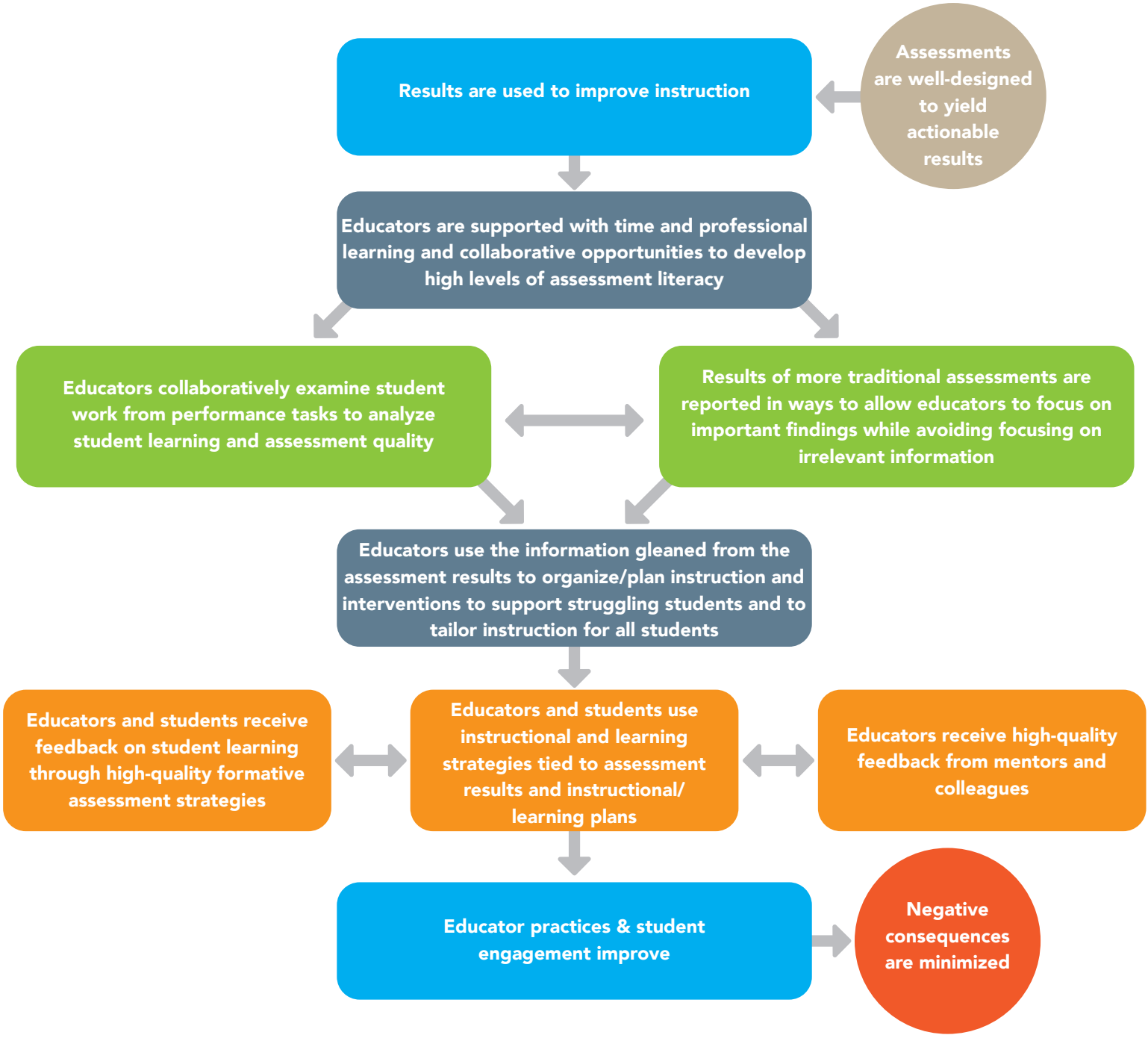


Figure 2. Expanded view of the theory of action.

As can be seen in Figure 2, many more details are needed to fully flesh out a theory of action. Once the state leaders and partners in the pilot come to agreement on an expanded theory of action, the project leaders would develop an implementation plan to guide the pilot activities. Additionally, and related most closely to the focus of this brief, the theory of action should be used to guide the formative evaluation so that the pilot can be supported by a continuous improvement process. We discuss below how one might use the theory of action to guide the formative evaluation. We draw on hypothetical examples from the figures below as well as real examples from New Hampshire's current innovative pilot operating under a waiver from ESSA, the Performance Assessment of Competency Education (PACE).

Starting from the Initial Indicators and Processes

Assuming the theory of action articulates an expected and intended chronology, the evaluation should begin with an examination of the most proximal, or near-term, indicators and processes. For example, based on the theory of action outlined in Figure 2, we expect that "educators are supported with time and professional learning and collaborative opportunities to develop high levels of assessment literacy" would occur before "educators use the information gleaned from the assessment results to organize/plan instruction and interventions to support struggling students and to tailor instruction for all students." Of course, these indicators are not perfectly sequential and we expect the system to iterate but to do so in a progressive manner. In this example, leaders must recognize that for educators to use information to better organize instruction, they must be provided with time and opportunities to improve their assessment literacy. Rather than throwing up one's hands and saying that "the system isn't working because educators are not using high quality instructional practices," a well-developed formative evaluation plan based on a theory of action would direct leaders to examine the quality of professional learning opportunities to ensure that educators have the tools necessary to improve their instructional practices.

This evaluative mindset requires us to ask key questions, such as those listed below, related to the design and implementation of the system for seeing progress on our proximal indicators:

Design-Related Questions	Implementation-Related Questions
• Does the design of the programmatic aspects of the assessment system effectively support change? • Is there a research base to support such design elements? • Do educators and key stakeholders have enough time to learn and internalize the intended changes in their roles and responsibilities? • Are there barriers that need to be removed from the existing system to support the intended changes?	• Are the intended programmatic aspects of the assessment system being implemented? • How do we know? • Are there different types of practices or conditions observed where we think implementation is "good" compared with "not as good?"

We cannot answer these questions unless we systematically collect data on these near-term indicators and processes. An example from New Hampshire's PACE program illustrates the importance of data collection and feedback.

The PACE theory of action posits that teachers will learn to score performance tasks consistently and accurately within and across districts. Processes and mechanisms associated with this indicator describe the need for high quality training and calibration protocols, opportunities for teachers to score performance task collaboratively (as part of calibration), and robust data collection and analytic procedures necessary for evaluating interrater consistency within each participating school district and cross-district calibration (accuracy).

Data from the first two years of PACE indicate very high degrees of interrater consistency within each participating district. However, in addition to within-district consistency, cross-district consistency or calibration is a critical aspect of the PACE innovation. Cross-district calibration is an evaluation of the degree to which educators from different districts score the same student work consistently. Cross-district calibration data are collected during a summer workshop that involves having educators from various districts look at student work from outside their districts to produce "consensus scores" used to evaluate scoring accuracy. The data provided from this calibration allows us to evaluate the accuracy component of the indicator described above. Districts must provide their teachers with high quality training and practice if they are to meet this indicator. Collecting these data is important, but providing feedback to districts for improving their local practices and therefore, the system as a whole, is what is most important.

While the cross-district calibration results varied considerably by grade span, subject area, and district, the first-year results revealed that the elementary teachers in one district were considerably more lenient (less rigorous) than teachers in other districts. These analyses were presented to the district's administrative team. They used this information to focus their scorer training in the following school year on ensuring that all teachers had internalized the expectations spelled out in the rubric and then had an opportunity to calibrate these expectations with other teachers in their school and district. The second-year results indicated that the teacher scoring in this district was consistent with the PACE teachers from other districts. In this case, we see how the theory of action and associated data collection for proximal indicators led to feedback and adjustments in order to improve the system.

Intermediate Indicators and Processes

The approach for evaluating intermediate indicators and mechanisms follows the approach described above for proximal indicators. The examination of intermediate indicators does not have to wait until the evaluation of proximal indicators is complete. Depending on the theory of action, both could happen simultaneously.

Again, an example from PACE may help illuminate the process. The PACE theory of action argues that through the work of developing, administering, and scoring the common tasks,[3] educators will learn to develop high-quality local tasks aligned to specific learning goals. We are using a multi-tiered approach to enhance and evaluate the quality of local assessments. The following elements are key aspects of this system:

- ✔ Investing in teacher-leaders to build local assessment expertise,
- ✔ Providing consistent and clear documentation and training materials on quality task development,
- ✔ Providing extensive professional learning opportunities for participating educators,
- ✔ Reviewing districts' "assessment maps" that describe the assessment coverage of the grade level standards and competencies,
- ✔ Reviewing a sample of local performance tasks from each participating district, and
- ✔ Reviewing student work samples as part of a "body of work" evaluation associated with the production of annual proficiency determinations.

It is beyond the scope of this brief to go into detail on each of these elements, but the important point is that a wealth of data is collected about both the design and implementation of a critical aspect of the PACE theory of action that is then used to provide targeted feedback to the state (system designers) and the district leaders (system implementers) in order to improve the quality of local performance tasks.

[3]A collaboratively designed "common task" is administered once in each grade and subject (math, ELA, and science) except for the three grades where the state standardized test is administered.

Distal Indicators and Intended Outcomes

We move finally to the evaluation of our most distal indicator(s), which in many cases provides information that may be too late and too coarse to inform programmatic changes without a well-designed backdrop of data collection related to the proximal and intermediate indicators. As noted previously, the first two guiding questions are addressed by the evaluation steps discussed up to this point. Examination of the long-term indicators/outcomes brings us to our third question in the set of three: Is the innovative assessment system producing the desired outcomes?

Using the example from Figure 1 on page 9, our distal outcome in the example theory of action is "student learning improves." Focusing on this distal outcome may seem like it allows us to efficiently answer our questions related to pilot effectiveness, but deeply understanding whether or not and how the system is working for the purposes of continuous improvement requires moving through the steps outlined to this point. The intended outcome of "student learning improves" will need more specificity to fairly evaluate whether in fact learning has improved. This would be a fairly common intended outcome for many innovative assessment system pilots, but we could easily imagine related outcomes that state leaders would specify such as, "students learn at noticeably deeper levels than previously" or "more students are truly college and/or career ready as a result of the pilot." No matter which of these is our intended outcome or outcomes, designers and evaluators need to be clear about how such outcomes will be measured and evaluated.

The evaluation of the PACE proximal indicators and processes was very internally-focused in that the project was not looking to make comparisons to non-pilot schools, but rather was hoping to see effective design features that lead to changes in practices, knowledge, skills, and dispositions among participating educators and others. While there is clearly some causal inference at play—i.e., did the innovative pilot cause these intended changes—the formative nature of the evaluation of the more proximal indicators allows for a little wiggle room in gathering the necessary data to support strictly causal inferences. However, when one hopes to attribute such things as increases in student learning or improvements in college and career readiness rates to the innovative pilot, we must contend with the considerable challenges of supporting causal claims in non-experimental settings. For example, if college and career readiness rates improve among students in the innovative pilot, how can leaders be sure this improvement is due to the innovative pilot rather than some other statewide or local initiative?

Randomly selecting students and schools to participate in the study and then randomly assigning schools (districts) to the pilot or non-pilot condition is the gold standard in social science research, yet it is often impractical, if not impossible. That said, leaders and evaluators must apply the most robust methods practical to evaluate the role of the pilot in influencing the intended outcomes.

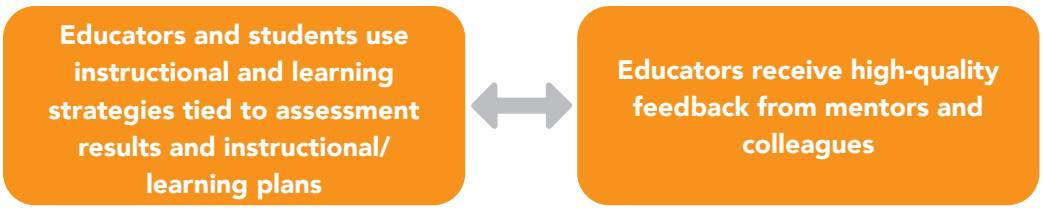
In addition to applying rigorous evaluation methods, it is important to consider the data sources that will be used to inform the evaluation of intended outcomes, especially outcomes related to improvements in student learning or student

achievement. Intuitively, one might think that the state standardized test scores should be the criterion for judging the quality of student achievement. States must consider statewide assessment performance when evaluating the success of their pilot. However, innovative assessment systems are often designed to promote and measure student learning in deeper and more meaningful ways and as such, states may find that student learning is underrepresented by state standardized assessment results alone.

States must also consider the requirement that the innovative assessment pilot expand to increasingly more school districts over time and eventually scale statewide within five years (or seven years if the state receives a two-year extension). Once a critical mass of districts is participating in the pilot, one must ask if it makes sense to rely heavily on the statewide standardized test as a criterion for evaluating student learning or should the state consider different measures of learning that better represent the type of outcomes the pilot is trying to promote.

Unintended Negative Consequences

Unintended consequences happen. Leaders and evaluators must attend to the potential for unintended negative consequences because there are almost always such consequences with any policy initiative, especially innovative reforms. It is critical for any formative evaluation to search for potential negative consequences of the reform so that they can be uncovered and addressed before they derail the pilot. As with evaluating the system design, implementation, and outcomes, the examination of possible unintended negative consequences can be drawn from the theory of action. A well-articulated theory of action clearly outlines the assumptions that must be upheld in order for the system to function as necessary, and a violation of those assumptions might result in a break-down of system efficacy, or worse, negative consequences for teachers and students. The assumptions of the system are represented by the arrows connecting the indicators in Figure 2, and can be articulated across the logical chain for all proximal (near-term), intermediate, and distal (long-term) indicators. Take for example the following two proximal indicators from Figure 2:

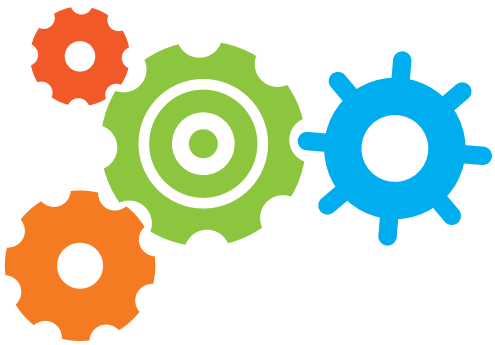


One of the assumptions necessary for this feedback system to work is that educators have adopted a growth mindset and are open to applying new instructional strategies in their own classrooms. If educators do not feel supported within a culture of innovation and growth, or have some other resistance to change, the assessment system may not function as intended, and implementation and expansion could lead to negative unintended consequences such as push-back from teachers, possible contention from educator unions, and lasting damage to the school culture.

An examination of the assumptions that underlie the theory of action is useful for detecting and preventing unintended negative consequences during any formative evaluation of the innovative assessment system. How does this fit into our guiding framework of the set of three questions? When the answer to any one question is yes, but the answer to any subsequent question is no, it is wise to evaluate what assumptions may have been violated and how those violations may be ameliorated. This ongoing process of continuous evaluation and improvement will help to prevent a possible cascade of unintended negative consequences associated with the innovative assessment pilot.

Summary

In coming full circle, we have tried to illustrate how the theory of action developed to guide the design of the innovative pilot should also be used to guide the formative evaluation as well. Such an evaluation supports a continuous improvement mindset—something many pilots are trying to instill in students—and recognizes that it is hard to get this "right" immediately out of the gate. Such an approach will allow us to steer towards ever increasing usefulness or utility.



Additional Support

KnowledgeWorks and the Center for Assessment are available to help states as they explore, design, and implement next generation assessment systems. Contact information for our organizations is listed below.

KnowledgeWorks can help states, districts, and other interested stakeholders establish the policy environments to support personalized learning at scale. The organization's expertise spans the federal, state, and district levels, supporting states with strategies to leverage current policy opportunities, remove existing policy barriers, and develop new policies that will help states create an aligned policy environment to support personalized learning. To learn more, contact the following people:

For State Policy and Alignment:
Matt Williams
Vice President of Policy and Advocacy
Williamsm@knowledgeWorks.org

For Federal Policy and Alignment:
Lillian Pace
Senior Director of National Policy
pacel@knowledgeworks.org

The **Center for Assessment** strives to increase student learning through more meaningful educational assessment and accountability practices. We engage in deep partnerships with state and district education leaders to design, implement, and evaluate assessment and accountability policies and programs. We strive to design technically sound policy solutions to support important educational goals. The Center for Assessment's professionals have deep expertise in educational measurement, assessment, and accountability and have applied this expertise to assessment challenges ranging from improving the quality of classroom assessments to ensuring the technical quality of state's large-scale achievement tests and ultimately to designing coherent assessment and accountability systems.

For Assessment and Accountability System Design and Strategic Implementation:
Scott Marion, Ph.D.
Executive Director
smarion@nciea.org

For Technical Quality and Comparability Design and Analyses:
Susan Lyons, Ph.D.
Associate
slyons@nciea.org

For Assessment Quality and Performance Assessment Development:
Jeri Thompson, Ed.D.
Senior Associate
jthompson@nciea.org

About Us

KnowledgeWorks is a national organization committed to providing every learner with meaningful personalized learning experiences that ensure success in college, career and civic life. With a presence in more than 30 states, we develop the capabilities of educators to implement and sustain competency-based and early college schools, partner with federal, state and district leaders to remove policy barriers that inhibit the growth of personalized learning and provide national thought leadership around the future of learning. www.knowledgeworks.org

The National Center for the Improvement of Educational Assessment, Inc. (Center for Assessment) is a Dover, NH based not-for-profit (501(c)(3)) corporation that seeks to improve the educational achievement of students by promoting enhanced practices in educational assessment and accountability. The Center for Assessment does this by providing services directly to states, school districts, and other organizations regarding the design, implementation, and evaluation of assessment and accountability systems. As a non-profit organization committed to the improvement of student learning, the Center for Assessment maintains a strong "open-source" ethic in terms of distributing its many creations and inventions. For example, the Center has developed many tools related to alignment methodology, student growth analyses, student learning objectives, comparability methods for innovative assessment systems, and validity evaluation that it provides freely to its clients and other non-commercial entities. www.nciea.org

The Nellie Mae Education Foundation is the largest philanthropic organization in New England that focuses exclusively on education. The Foundation supports the promotion and integration of student-centered approaches to learning at the middle and high school levels across New England—where learning is personalized; learning is competency-based; learning takes place anytime, anywhere; and students exert ownership over their own learning. To elevate student-centered approaches, the Foundation utilizes a four-part strategy that focuses on: building educator ownership, understanding and capacity; advancing quality and rigor of SCL practices; developing effective systems designs; and building public understanding and demand. Since 1998, the Foundation has distributed over $180 million in grants. For more information about the Nellie Mae Education Foundation, visit www.nmefoundation.org.