RAND CORPORATION

Research Report

JULIA H. KAUFMAN, ELIZABETH D. STEINER, JONATHAN SCHWEIG, SOPHIE MEYERS, KAREN CHRISTIANSON

Evidence on the Validity, Reliability, and Usability of the Measuring and Improving Student-Centered Learning (MISCL) Toolkit

Key Takeaways and Lessons on Developing Tools for School Improvement

Sponsored by the Nellie Mae Education Foundation

KEY FINDINGS

- Our research suggests that the Measuring and Improving Student-Centered Learning (MISCL) Toolkit instruments measured student-centered learning (SCL) constructs as intended and might differentiate among schools with differing levels of SCL.
- Some but not all of the items in the SCL survey scales appeared to be closely related and measured the same constructs.
- Relationships of SCL with external variables, such as achievement, were not always consistent with theory.
- Users found the MISCL Toolkit process understandable and useful, although concerns about the burden of administering the MISCL Toolkit and its potentially evaluative nature led to further revisions of its content.

For decades, student-centered learning (SCL) has been a buzzword for a variety of approaches that keep students' goals, interests, and needs central to the teaching and learning process. The terms associated with SCL include *personalization*, *differentiation*, *problem-based learning*, *competency-based learning*, and *inquiry-based learning*, among many others. Despite the proliferation of SCL approaches, or perhaps because of this proliferation, researchers and practitioners are still learning about which SCL strategies are the most supportive of students and what school-level contextual conditions and resources are needed to support high-quality SCL.

Abbreviations

ATP	American Teacher Panel
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
FRL	free or reduced-priced lunch
MISCL	Measuring and Improving Student-Centered Learning
NMEF	Nellie Mae Education Foundation
PSAT	Preliminary Scholastic Aptitude Test
RMSEA	Root Mean Square Error of Approximation
SCL	student-centered learning
SRMR	Standardized Root Mean Residual

To support high schools seeking to measure, understand, and reflect on SCL, the Nellie Mae Education Foundation (NMEF) has focused on four key principles of SCL:

1. Learning is personalized.
2. Learning is competency-based.
3. Learning occurs anytime, anywhere.
4. Learning is student-owned.

NMEF provided support for the RAND Corporation to design and test the validity, reliability, and usability of a toolkit to help high schools measure and reflect on SCL, as defined mainly through these four principles. In this report, we share some results from our validity, reliability, and usability testing and some lessons for development of similar toolkits. In particular, such toolkits are intended to help *school practitioners*, and not researchers, examine teaching and learning. Therefore, those developing toolkits may wish to keep in mind a set of design principles that maximize the potential for useful data and minimize burdens and common issues related to collecting, analyzing, and reflecting on the data.

We begin this report by providing an overview of the research on SCL to date that points to the need for user-friendly tools to help schools measure and reflect on SCL. Then, we provide an overview of the Measuring and Improving Student-Centered Learning (MISCL) Toolkit and how we approached its testing. Next, we provide an overview of our methods—the types of evidence we collected and analyses we conducted. We then discuss our results, which include a description of findings for each section of

our Toolkit and how we revised the Toolkit's content in response to what we learned. We conclude our report with a brief summary of the key findings and implications for researchers seeking to develop and practitioners considering use of similar toolkits. The Toolkit is available, along with some resources to support administration and analysis of the surveys that are part of the Toolkit, at studentsatthecenterhub.org/resource/measuring-scl-toolkit/.

What Do We Know About SCL?

Prior to developing the MISCL Toolkit, we conducted a literature review that focused on defining SCL and the relationship between SCL strategies and student outcomes, adhering closely to the principles that NMEF used to define SCL. The literature review is available online at studentsatthecenterhub.org/resource/measuring-scl-toolkit/. We found that the effects of SCL interventions on student engagement and achievement were largely mixed. Research on some SCL interventions has tied particular strategies to improved student outcomes. Such strategies include personalization of content based on student interests (Walkington, 2013; Ku and Sullivan, 2002; Awofala, 2016); competency-based or mastery-based systems that allow students to work at their own pace (Kulik, Kulik, and Cohen, 1979; Kulik, Kulik, and Bangert-Drowns, 1990; Abakpa and Iji, 2011; Friedlaender et al., 2014); service-learning programs (i.e., programs focused on involving students in experiences that benefit others or their communities) with strong linkages to curriculum (Furco, 1996; Billig, 2000; Billig, 2009; Conway, Amel, and Gerwien, 2009; Celio, Durlak, and Dymnicki, 2011); and teaching students megacognitive strategies (Cross and Paris, 1988; Cardelle-Elawar, 1992; Boulware-Gooden et al., 2007). That said, there are also cases in which use of these interventions and strategies showed no evidence of improved outcomes or less evidence for some of these strategies relative to others (e.g., Steele et al., 2014; Lopez and Sullivan, 1992; Cole, Kemple, and Segeritz, 2012; Cavanaugh et al., 2004). Existing research also suggests that SCL is challenging for educators to implement as intended, given that it requires them to attend closely to the needs of all students in their classrooms (Pane et al., 2017; Steiner et al., undated). Some studies also suggest that allowing students the autonomy to vary the pace and the focus of their individual instruction could increase

How Was SCL Defined and Measured Through the MISCL Toolkit?

The MISCL Toolkit defined *SCL* broadly as follows:

SCL is students' deep engagement in learning opportunities that are designed to address their goals and interests while at the same time providing appropriate supports and challenges according to their learning needs.

The strategies specifically measured through the MISCL Toolkit drew heavily from NMEF's SCL principles:

- Strategy 1. Learning is personalized to align with students' needs, interests, and pace.
- Strategy 2. Learning is challenging and engaging and meets students at the point where they are in a competency-based system.
- Strategy 3. Learning can happen anytime, anywhere.
- Strategy 4. Learning opportunities promote student agency and ownership.
- Strategy 5. Learning is informed by data.

The MISCL Toolkit instruments also gathered information about the contextual conditions that might support SCL, including systems for continuous improvement; people, policies, and infrastructure that support SCL; and learning environments. The Toolkit is available online at studentsatthecenterhub.org/resource/measuring-scl-toolkit/.

Given the rapid growth and embrace of SCL approaches nationwide, there is a demand for clearer definitions of SCL.

achievement gaps and lead to inequitable access to learning opportunities (Steele et al., 2014).

Despite limited research and apparent challenges to the provision of high-quality SCL approaches, increasing numbers of states and school systems are pursuing interventions that include aspects of SCL. Numerous school systems across the United States have been experimenting with personalized and competency-based student learning approaches (Gross and DeArmond, 2018; Pane et al., 2017; Steiner et al., 2017). In addition, as of 2016, ten U.S. states either had policies or played an active role in supporting students to pursue "competency-based pathways through credit flexibility" that allow students to master knowledge and skills at their own pace (Patrick et al., 2016), and the majority of other U.S. states were exploring such policies.

Given the rapid growth and embrace of SCL approaches nationwide, there is a demand for clearer definitions of SCL, including better ways to collect data on the extent of SCL in classrooms, track SCL progress over time, and support improvement of SCL. In our own recent review, we identified nearly 100 existing instruments intended to measure some aspect of SCL, including surveys, interview or focus group protocols, and practice guides for school systems (Steiner et al., undated). These instruments typically define and measure SCL in various ways but often use only one method for data collection and measurement of SCL (e.g., a survey or a practice guide) and typically only collect data from one—or occasionally two—data sources, such as instructional staff or students. In addition, most of these instruments, with a few exceptions, provide schools with the opportunity to collect data but do not support them to interpret the data or engage in a process to reflect on and improve SCL in their own settings.

What Is the MISCL Toolkit?

The MISCL Toolkit was developed to help school practitioners or other stakeholders—such as an external professional development provider who would like to gather baseline data to inform the design of professional learning activities—measure, understand, and reflect on the extent of SCL in high schools. It was designed to enable data collection on SCL and supports for SCL present in high schools through a variety of measures and from those in several roles (from district and school leaders and instructional staff to students).

There are many ways to define and think about SCL. We developed the MISCL Toolkit to gather information on SCL related closely to NMEF's SCL principles: Learning is personalized; competency-based; occurs anytime, anywhere; and is engaging. In addition, drawing on the SCL literature and other related work, we also developed the MISCL Toolkit to measure the extent to which learning was informed by data and to capture elements of contextual conditions that might support SCL.

Overall, the MISCL Toolkit consists of a User Guide, six instruments for collecting data about SCL, and Support Tools for helping users reflect on the data. The instruments themselves are intended to collect evidence on SCL practices within high schools and across multiple individuals in different roles so that users can construct a holistic understanding of SCL implementation in their schools. The User Guide includes instructions on how to administer the instruments, analyze the data collected, discuss next steps, and communicate SCL continuous improvement efforts. Finally, the Support Tools include (1) a Quick Start Guide that includes an overview of the Toolkit and how to start; and (2) a Reflection Conversation Guide to support a conversation about the data collected, and how to use that data to consider the next steps for improving SCL.

MISCL User Guide

The User Guide walks users through the process of measuring and improving on SCL in their contexts. The steps of this process, which are outlined in the User Guide, are

- set an SCL improvement focus
- plan administration and collect data (through surveys, a student focus group, and a classroom walkthrough)
- explore the data and plan next steps
- communicate findings for continuous improvement.

The MISCL Toolkit is intended to be "self-serve," so that practitioners can use the User Guide to help guide them through the Toolkit process.

MISCL Instruments

The MISCL instruments included surveys of district and high school leaders, instructional staff (i.e., teachers and other educators supporting student learning in schools), and students; a student focus group; and a Classroom Walkthrough Guide.

- **The MISCL Toolkit surveys** include some parallel items to compare the different perspectives among district/school leaders, instructional staff, and students. As noted in Table 1, each survey includes at least one question on each of the five key SCL strategies measured through the Toolkit, along with items on the contextual conditions that support SCL.
- **The student focus group** is intended to gather rich examples of aspects of SCL and supports for SCL in the high school, from students' perspectives, as a supplement to what is gathered through the surveys.
- **The Classroom Walkthrough Guide** is intended to help users collect "snapshots" of how SCL is taking place inside classrooms, related to three SCL strategies that can be readily observed: personalization, competency-based learning, and student ownership and agency over learning. Prior to conducting the walkthrough, participants identify several key aspects of SCL to observe, drawing from the example "look-fors" provided in the Walkthrough Guide. The participants then rotate to three to four classrooms to note the presence of each look-for they identified.

Table 2 provides additional information about the specific aspects of SCL that are measured through the instruments for each strategy. We developed these aspects of SCL through discussions with NMEF on SCL, our literature review findings, and feedback from our seven-member Advisory Board. The board included practitioners (e.g., teachers, principals, technical support providers) and researchers with expertise in studying, measuring, or implementing SCL. Each aspect is measured through at least one item or a set of items within each instrument, with the exception of the Walkthrough Guide, which only focuses on aspects of SCL that can be observed directly in a classroom.

MISCL Support Tools

Support tools for the MISCL Toolkit include the following:

- **Quick-Start Guide.** This tool provides snapshots of the Toolkit process and key considerations for anyone thinking about using the Toolkit. The Quick-Start Guide includes a Planning Worksheet that provides guidance to users on how to administer the Toolkit and document their progress as the different steps in the process are completed, and a suggested timeline for completing all steps in the process.
- **Reflection Conversation Guide.** This tool is designed to help facilitate a conversation between various stakeholders (e.g., school leaders, instructional staff, and students) about the SCL data collected from the MISCL instruments and how to improve SCL practices within a school. In this conversation, participants are encouraged to compare data across all of the instruments and discuss differences they see in SCL practices by different subgroups (e.g., subjects, grade levels).

Users are strongly encouraged to administer the Toolkit in full. However, if a school or district is interested in getting a quick read from a few groups

TABLE 1

MISCL Instruments

MISCL Instrument and Description	Area of SCL Emphasized: Personalization	Competency-Based Learning	Anytime/ Anywhere Learning	Student Agency and Ownership	Data Use	Contextual Conditions Supporting SCL
District leader survey: 20-minute survey to district leaders						
School leader survey: 25-minute survey to high school leaders						
Instructional staff survey: 30-minute survey to high school teachers						
Student survey: 20-minute survey to high school students						
Student focus group: 60-minute focus group with six to eight high school students						
Walkthrough: Guide for walking through several classrooms to observe SCL in action						

NOTE: Green shading = addressed comprehensively through several measures. Yellow shading = addressed to some degree through one to two measures. Gray shading = not addressed.

TABLE 2

SCL Strategies and Aspects of Each Strategy Measured Through Toolkit Instruments

SCL Strategy	Aspects of Each SCL Strategy Measured Through Toolkit Instruments
Personalization	• Educators and students work together to personalize students' pathways through content and courses • Timing and delivery of learning opportunities are varied to support students' learning needs, interests, and pace • Assessments are varied to support students' learning needs, interests, and pace
Competency-based learning	• Learning targets and pathways are clear, measurable, and competency-based • Courses, assignments, activities, and assessments are aligned to competencies • Students access assessments when they are ready to demonstrate mastery • Learning opportunities and assessments reflect high expectations and provide appropriate challenge for each student • Students engage in meaningful, cognitively challenging assignments and activities
Anytime/anywhere learning	• Students engage in multiple credit-bearing learning activities within and outside the classroom • Students engage in authentic assessments and activities with connections to the real world
Student agency	• Students participate in activities that promote self-regulation, collaboration, metacognition, and communication strategies • Students develop their own learning pathways and profiles with appropriate support
Data use	• Educators and students gather data on students' needs, interests, goals, and learning progress • Educators and students use data to inform learning pathways and monitor progress
Contextual Conditions	• School systems have systems in place for continuous improvement to support SCL • The people, policies, and infrastructure within a school system support SCL • The learning environments within a high school support SCL

of respondents or testing out the Toolkit before committing to full use, the Toolkit is designed for selected instruments to be administered to a more limited group of participants. If possible, at least two instruments should be administered (e.g., the student survey and instructional staff survey) to ensure that perspectives from multiple stakeholders are captured.

MISCL Toolkit Development and Testing Overview

The MISCL Toolkit was developed, tested, and revised through an iterative process that included a literature review, advisory board vetting, two pilot tests, and administration of the instructional staff and leader surveys to large survey samples. Table 3 provides an outline of the timeline for our development and testing process.

The first pilot test was conducted in two northeastern United States school districts and focused on testing the survey instruments. After revisions in response to that pilot test, we did a second pilot test of the entire MISCL Toolkit in two additional northeastern U.S. school systems (including one charter school system and one traditional public school district).

This second pilot test included all of the instruments, Support Tools, and usability testing. To do that testing, we observed the entire MISCL Toolkit process, including how users within each school system examined the collected data and had reflection conversations about the data. We interviewed respondents about their experiences with all of our instruments to understand their response processes and the extent to which the data collected through the MISCL Toolkit process were useful and informative.

We also collected survey data from two additional samples to supplement the smaller number of surveys we were able to collect through our pilot tests. The two additional samples were

- a national sample of high school teachers through the RAND American Teacher Panel (ATP)

TABLE 3
Timeline for Development and Testing of MISCL Toolkit

Date	Activity	Type of Evidence Collected
January–August 2017	• Conducted literature review • Developed framework for measuring and improving SCL and MISCL instruments • Vetted framework and instruments with Advisory Board	• Validity evidence related to content for SCL framework and survey instruments
November 2017	• Pilot-tested instruments (Pilot Test I)	• Validity evidence related to response processes for survey instruments
December 2017–April 2018	• Revised MISCL instruments based on feedback from Pilot Test I • Created MISCL User Guide and Support Tools	
September 2018–March 2019	• Pilot-tested entire MISCL Toolkit (Pilot Test II) • Fielded MISCL surveys to supplemental sample of teachers, school leaders, and district leaders	• Validity evidence related to response processes, internal structure, and external variables for instruments • Reliability for instruments • Usability of MISCL Toolkit
April–September 2019	• Analyzed usability, validity, and reliability evidence	
October–December 2019	• Revised MISCL Toolkit	

In keeping with the intended purpose of the MISCL Toolkit, we collected a range of validity evidence aligned with the purpose of the MISCL as a tool for reflection and continuous improvement.

We did not collect evidence regarding the validity of the MISCL Toolkit for accreditation, evaluation, or other high-stakes uses.

We also collected evidence on the Toolkit's usability—whether school staff were able to use the MISCL User Guide and Support Tools as intended.

- a convenience sample of district leaders and high school principals in five New England states.

The types of evidence collected during the testing of the MISCL Toolkit are described in more detail in the next section.

Data and Methods

The intended purpose of the MISCL Toolkit informed the types of validity, reliability, and usability evidence we collected and the contexts in which we collected such information. As we described earlier in this report, the MISCL Toolkit aims to help school practitioners (and other stakeholders) measure, understand, and reflect on the extent of SCL in their high schools with the ultimate goal of improving implementation. It is not intended to be used for accreditation, evaluation, or other high-stakes purposes. Our analyses were designed with two main questions in mind:

1. **Validity and reliability:** To what extent did the MISCL instruments consistently and precisely measure what they were intended to measure, drawing on evidence related to content, response processes, internal structure, relationships with external variables, and reliability?
2. **Usability:** To what extent was the MISCL Toolkit usable and useful to those who undertook the Toolkit process?

To answer these questions, we collected a variety of evidence. In particular, we did the following:

- To collect evidence on the validity and reliability of MISCL instruments, we conducted a literature review of areas to be measured by the instruments; gathered advisory board feedback on the instruments; and collected data through the instruments from samples of school and district leaders, instructional staff, and students. We also conducted cognitive interviews with a small number of users in each sample to better understand how they interpreted survey items (i.e., response processes).
- To collect evidence on the usability of the MISCL User Guide and Support Tools, we observed two school systems engaging in the entire MISCL Toolkit process, including setting goals for administration, fielding instruments, analyzing and reflecting on the data, and planning next steps to improve SCL.

In the remainder of this section, we present an overview of the types of data we collected, how we collected such data, and how we analyzed the data related to these questions. We then describe our data sources and analysis in more detail. Additional details on the data sources and analytic methods can be found in Appendix A.[1]

Types of Evidence Collected

To assess whether the MISCL instruments measured what they were intended to measure—the extent of SCL implementation in high schools—we collected the following types of **validity evidence**, consistent with established guidelines provided in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014):

- evidence that the content of the instruments faithfully and fully represents the constructs

they are intended to measure—we refer to this as **evidence based on content**
- evidence that those who complete the instruments are interpreting items in the same way and consistent with developer intentions—we refer to this as **evidence based on response processes**
- evidence that the items in each survey scale measure one underlying SCL construct and that survey scales correlate in anticipated ways (e.g., whether items intended to measure student engagement correlate strongly with one another and whether scale scores for student engagement and personalization are positively correlated)—we refer to this as **evidence based on internal structure**
- evidence that scales in our instruments are associated with external measures (e.g., student achievement, student characteristics, school characteristics) in ways that are supported by research-based evidence or a well-founded theoretical justification—we refer to this as **evidence based on relationships with external variables.**

We gathered all of the above types of validity evidence for the MISCL survey instruments. For the student focus group, we collected information about response processes only. For the Walkthrough Guide, we gathered some evidence on response processes, but that information was mainly focused on its usability. For this reason, we therefore examined usability of the Walkthrough Guide as part of the entire MISCL Toolkit but did not examine its validity and reliability in the same way that we did for the other instruments.

Following the *Standards for Educational and Psychological Testing*, we examined the **reliability** of the scale scores derived from the survey instruments and the extent to which there was evidence that these scores were consistent, precise, and not unduly influenced by random measurement error (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

In this investigation, we took two approaches to quantifying reliability. First, we examined the extent to which survey respondents responded similarly across all of the items in a particular scale. This provided information about the extent to which students, instructional staff, school leaders, and district leaders were distinguishable from one another based on their survey responses. Second, we examined the extent to which respondents in the same school who were making judgments about the same phenomena agreed with one another in their appraisals. This provided information about the extent to which survey responses were consistent. Although there are other ways to quantify reliability (e.g., the precision of classroom-level and school-level scores, the extent to which classrooms and schools can be distinguished based on the average scores of respondents), it was not possible to use those approaches with the current data. Nonetheless, the investigations presented here provided important preliminary evidence about reliability.

We assessed whether the MISCL Toolkit was used as intended by observing and gathering data on the **usability** of the entire Toolkit at two high schools in different districts with different levels of SCL implementation. Using principles of user-centered design, we defined *usability* in terms of whether the MISCL Toolkit was regarded as both usable and useful by users (Dumas and Redish, 1993; Gould and Lewis, 1985). Over the course of two-day site visits, we observed district and school leaders, instructional staff, and students using the User Guide, Support Tools, and Walkthrough Guide and conducted debriefing conversations with them about their experience using the different components of the MISCL Toolkit. The goals of our observations and conversations were to determine whether the User Guide, Support Tools, and MISCL instruments were used as intended, to identify any problems participants encountered when using them, and to assess their satisfaction with the User Guide and Support Tools.

Data Sources

We relied on various sources of data to investigate the validity, reliability, and usability of the Toolkit. Here, we outline the different samples used in our analysis. Appendix A includes additional information about the roles and numbers of those who

participated in interviews and focus groups as part of the pilot tests and details on response rates and survey administration.

In addition to the data sources listed below, we relied on a comprehensive literature review of SCL and expert feedback to inform the development of our SCL framework, instruments we used to measure SCL, and the MISCL Toolkit. Our literature review results are summarized in a separate publication,[2] and an overview of the literature review and advisory board feedback process is included in Appendix A.

Pilot Test I sample. We fielded the surveys in two school districts as part of Pilot Test I, and we conducted cognitive interviews with survey respondents to determine how well they understood the survey items and could respond to them. The two schools that participated in Pilot Test I were chosen because of their contrasting contexts and experiences with SCL. See Table 4 for an demographic overview of the two school districts.

Pilot Test II sample. After revising the surveys in response to Pilot Test I, we selected two additional school systems to participate in Pilot Test II, which involved fielding all Toolkit instruments and engaging in all steps outlined in the Toolkit User Guide, including setting goals for data collection, analyzing the data after administration of the instruments, and reflecting on the data. We gave these schools limited input on how to use the Toolkit within their contexts, because the Toolkit is intended to be "self-serve" for those who wish to use it. The two high schools within the two school systems that participated in Pilot Test II were chosen to differ in terms of both their contexts and experiences with SCL (see Table 4). We conducted two-day in-person visits to each Pilot Test II school to observe Toolkit use, including a classroom walkthrough, student focus group, and reflection conversation. Altogether, we conducted 46 cognitive interviews and focus groups with district and school leaders, instructional staff, and students that focused on how participants understood and responded to survey items and student focus group questions, how they felt during the walkthrough, how the reflection conversation went, and whether they had any suggestions for improvement. The results of this analysis were used to inform revisions to the survey instruments, Support Tools, and User Guide.

District and school leader survey sample. The district and school leader survey sample included both the small numbers of district and school leaders who completed surveys as part of Pilot Test II (n = 10) and a convenience sample of district and school leaders in five New England states—Maine, New Hampshire, Vermont, Connecticut, and Rhode Island—including 47 superintendents and 102 high school principals.

Instructional staff survey sample. The instructional staff survey sample included both the

TABLE 4

Characteristics of High Schools That Participated in Pilot I and Pilot II

	Pilot I, School I	Pilot I, School II	Pilot II, School I	Pilot II, School II
Number of students	1,500–1,600	100–200	400–500	200–300
Number of teachers	140–150	10–20	40–50	20–30
Urbanicity	Suburb	Rural	Suburb	City
Student ethnicity	0–10% Asian 20–30% black 20–30% Hispanic 30–40% white	0–10% Asian 0–10% black 0–10% Hispanic 90–100% white	0–10% Asian 0–10% black 0–10% Hispanic 80–90% white	0–10% Asian 20–30% black 50–60% Hispanic 10–20% white
Title I school?	No	Yes	No	Yes
Percentage of students receiving free or reduced-priced lunch (FRL)	70–80%	40–50%	10–20%	90–100%
Experience with SCL[a]	Beginning	Intermediate	Beginning	Experienced

[a] Experience with SCL was determined based on how the school described itself to us.

instructional staff who completed surveys as part of Pilot Test II (n = 37) and the RAND ATP sample (n = 553), a geographically diverse panel of teachers from across the United States. The full ATP consists of more than 24,000 teachers across the United States who have agreed to respond periodically to education-related surveys. Our randomly selected ATP subsample consisted of 553 high school teachers in 518 schools.

Student survey sample. The student survey sample included students who completed surveys as part of Pilot Test II (n = 377).

Analysis

In this section, we describe analysis we conducted based on validity, reliability, and usability evidence. Table 5 outlines the data we collected and analyses we conducted.

Validity evidence based on content. We developed the instruments and gathered evidence based on content by conducting a literature review and vetting the instruments with our Advisory Board. Our literature review included 156 studies and helped us identify the five key SCL strategies and underlying contextual conditions that we later used to construct the MISCL surveys. Our Advisory Board provided feedback on the wording of the SCL strategies and contextual conditions and the particular SCL practices included within each strategy. Our Advisory Board members also reviewed draft survey instruments and provided feedback on whether they believed survey items would appropriately measure the five SCL strategies and contextual conditions.

Validity evidence based on response processes. We analyzed cognitive interview data from administrators, instructional staff, and students we interviewed in Pilot Test I and Pilot Test II after those groups took our surveys and participated in the student focus groups. The cognitive interviews focused on how participants in each group understood and responded to survey items and focus group questions and also on their suggestions for improvement. The results were used to understand whether respondents were interpreting the questions in the same way and as intended, and to improve and refine the instruments and inform development of the User Guide.

Validity evidence based on internal structure. We analyzed MISCL survey results, using data from all samples, to assess the internal structure of survey items. We specifically merged Pilot Test II data with ATP data to examine item-level data from the instructional staff survey, and we merged Pilot Test II data with New England district and school leader sample data from the district and school leader surveys.

Given that relatively little is known about the specific SCL constructs we measured in our survey, we identified numerous scales across all of the surveys to better understand the extent to which our survey was measuring particular aspects of SCL. For example, to measure aspects of personalization in the student survey, scales on each of the following topics were examined: student choice, tailored learning opportunities, diverse learning opportunities, and personalization supports. Similarly, in the teacher survey, scales on the following topics were examined to measure personalization: student choice and personalized learning opportunities. Scales are described in more detail in Appendix B, which also includes a list of the items in each scale for each survey.[3]

Facilitators and participants in the MISCL Toolkit process might use scales to obtain feedback about key aspects of SCL in their schools. For example, a school with a strategic plan focused on improving data-informed decisionmaking might focus on the Data Use scales. Or, a school focusing on strengthening climate and culture may focus on the Contextual Conditions Supporting SCL scales.

We conducted a confirmatory factor analysis (CFA) on all the student, instructional staff, and school leader survey data we collected to confirm that sets of items could be combined to form scales as intended. We did not conduct a CFA for the district leader survey because the sample was too small to allow for us to do so. Each scale was evaluated separately.

We elected to use a CFA for these investigations because CFA provides a statistical framework that allows us to test whether items in a scale do, indeed, measure a single construct. CFA also provides diagnostic information about items that are not functioning as intended in specific scales. Specifically, we used four sources of diagnostic information to

TABLE 5

Types of Validity, Reliability, and Usability Evidence Collected and Analyses Conducted

Type of Evidence	Data Collected	*n*'s and Response Rates (if Applicable)	Analyses Conducted
Evidence based on **content**	• Literature review	• 156 studies included	• Identify key tenets of SCL and collect evidence on efficacy of SCL approaches
	• Feedback from an advisory board of experts	• Seven members	• Review of instruments by advisory board • Qualitative analysis of advisory board feedback to inform instrument revisions
Evidence based on **response processes**	• Pilot Test I (fielding of surveys only) in two school systems	• Interviews with nine students, ten instructional staff, and four school and district leaders	• Qualitative analysis to inform revisions to the MISCL Toolkit (from Pilot Test I and Pilot Test II)
	• Pilot Test II (use of the entire MISCL Toolkit, including all instruments and surveys, student focus group, and walkthrough) in two school systems	• Interviews with 18 students, 22 instructional staff, four school leaders, and two district leaders	
Evidence based on **internal structure**	• Pilot Test II survey data • Teacher survey; RAND ATP data • District and school leader surveys; Northeast sample of superintendents and principals	• Student survey (*n* = 377; overall response rate: 58%) • Instructional staff survey (*n* = 590; overall response rate: 52%) • School leader survey (*n* = 107; overall response rate: 12%) • District leader survey (*n* = 52; 7%)	• Confirmatory Factor Analysis (CFA) of student, teacher, and school leader surveys • Correlations of scale scores within survey instruments
Evidence based on **relationships with external variables**	• Pilot Test II student survey data, alongside student achievement and student/school demographic data (see Table 6) • School demographic data examined with ATP data (see Table 6) • Pilot Test II student and teacher survey data, compared with self-reported levels of SCL from school administrators	• Student survey (see *n*'s and response rates above) • ATP sample (*n* = 553, response rate: 63%)	• Regression with concurrently measured variables (student survey only) • Comparison of scale scores by subgroup
Evidence of **reliability/ precision**	• Pilot Test II survey data • Teacher survey; RAND ATP data • District and school leader surveys; Northeast sample of superintendents and principals	• Student survey; instructional staff survey; school leader survey; district leader survey (see above for n's and response rates)	• Internal consistency • Within-school consensus (student survey only)
Evidence of **usability**	• Pilot Test II observations of MISCL Toolkit use • Pilot Test II interviews with respondents and participants in reflection meetings on usability and usefulness of MISCL Toolkit	• Observations of use at two schools • Interviews with nine instructional staff and four school leaders	• Qualitative analysis of observational and interview data

NOTES: Response rates were calculated by dividing the total number of respondents to whom we reached out (sometimes across multiple samples) for each survey by the total number who responded. The samples were from the schools that participated in Pilot Test II, along with the ATP and New England school and district leader sample. Response rates to the New England school and district leader sample were low but in line with other similar past survey efforts across a large number of school leaders (e.g., Kaufman et al., 2015).

evaluate the integrity of the scales: a chi-square based test statistic, a Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Residual (SRMR).[4] Following established recommendations in the literature (Hu and Bentler, 1999; Steiger, 1990; Vandenberg and Lance, 2000), a nonsignificant chi-square test statistic, an $RMSEA \leq 0.06$, a $CFI \geq 0.95$, and an SRMR ≤ 0.05 were taken as suggesting strong evidence supporting our scales. $CFI \geq 0.90$, $RMSEA \leq 0.08$, and SRMR ≤ 0.08 were taken as suggesting adequate evidence supporting our scales. If a scale had weak supporting evidence, we examined the patterns of factor loadings, the model residuals, and the modification indexes to understand as best as possible the specific issues with the scales and to make modifications to the scales where possible to improve their integrity. If these modifications could not improve the quality of the scales, we noted that there was not adequate evidence of the internal structure to support the scales' use.

In almost all cases, we were able to identify specific issues with the scales that compromised their integrity. Often, the diagnostic evidence suggested that some pairs of items were correlated more strongly than could be explained by the fact that the items measured a single construct. In these cases, we identified items that were redundant in content and removed one item. In other cases, it was determined that although subsets of items overlapped considerably, they still captured something important about the construct. In these cases, we left the items in the survey but acknowledged the excess item overlap by allowing specific pairs of items to correlate with one another.

Finally, we created composite scale scores using simple averages representing each of the factors and calculated the correlations among these scale scores to determine the extent to which there was evidence that the scales were measuring distinct constructs that were associated in ways that were consistent with theory. For example, we would expect two scales measuring complementary SCL practices to correlate positively, and we would likewise expect scales measuring SCL practices to correlate negatively with scales about school conditions that inhibit SCL practice adoption.

Validity evidence based on relationships to external variables. Using our review of the existing literature, we developed a priori predictions about the strength and direction of the relationships among survey scales and measures of student achievement, student characteristics, and school characteristics. These hypotheses are summarized in Table 6. We hypothesized that student achievement would be positively related to the SCL constructs measured in our surveys. This hypothesis is informed by recent quasiexperimental studies showing that students in schools using personalized learning practices or SCL practices showed greater achievement growth than their peers (Pane et al., 2015; LaBanca et al., 2015). One caveat for this hypothesis is that many studies have not identified a clear positive relationship between SCL and achievement (e.g., Steele et al., 2014; Lopez and Sullivan, 1992; Cole, Kemple, and Segeritz, 2012; Cavanaugh et al., 2004), which suggests that SCL might be related to higher achievement in some settings but not others.

We also hypothesized that student discipline (including suspensions) would be negatively associated with SCL constructs. This hypothesis was informed by prior research demonstrating that student-centered environments play a key role in reducing problem behaviors (Austin, 1979; Doyle, 1990). We hypothesized that lower-income and minority students and those in low-achieving schools would engage in fewer informal learning opportunities outside school and face higher exposure to lecture-based, teacher-centered instruction (Lewis et al., 2014; Oakes et al., 1990). We also thought that younger students might report experiencing less SCL than their older peers, given reports by school administrators who typically expected students in later high school grades to be able to engage in SCL to a greater degree than incoming ninth graders. Finally, we hypothesized that schools that had more experience with SCL—according to their self reports—would have had higher ratings of their SCL practices because their SCL model had been more fully implemented.

We tested these hypotheses using the available data for each group of survey respondents, which were made available to us by the schools participating in Pilot Test II.[5] The only student achievement

TABLE 6
A Priori Hypothesis and Evidence

External Variable	Hypothesized Relationship	Data Collected	Citations
Achievement	Presence of SCL practices would be associated with higher achievement	PSAT scores from 11th and 12th grade students in Pilot Test II schools	Pane et al., 2015 LaBanca et al., 2015
Discipline	Presence of SCL practices would be associated with lower rates of suspension	Suspension data from students in Pilot Test II schools	Austin, 1979 Doyle, 1990
Income	Lower-income students would experience fewer SCL-aligned practices	Title I eligibility, teacher and school leader survey samples (Pilot Test II, ATP survey, and New England school leader survey respondents)	Lewis et al., 2014 Oakes et al., 1990
Race/ethnicity	Black and Hispanic students would experience fewer SCL-aligned practices	Race/ethnicity data from students in Pilot Test II schools Racial/ethnic composition of the school; teacher and school leader survey samples (Pilot Test II, ATP survey, and New England school leader survey respondents)	Lewis et al., 2014 Oakes et al., 1990
Student age	Younger students would report experiencing less SCL than their older peers	Grade level	Reports from school administrators
SCL implementation	Schools that self-reported higher levels of experience with SCL implementation would have higher SCL scale scores	School identifier for students in Pilot Test II schools	Reports from school administrators

NOTE: PSAT = Preliminary Scholastic Aptitude Test.

data available to us across Pilot Test II schools were Preliminary Scholastic Aptitude Test (PSAT) scores, which we used for our achievement analysis, despite the potential drawbacks of using a single achievement measure that might not be sensitive to the types of learning promoted through SCL. PSAT scores were available for the majority of 11th graders in both schools (approximately 95 percent) and for the majority of 12th graders in one school. We also had data on student characteristics (i.e., race/ethnicity, FRL, grade level) and discipline (i.e., in-school and out-of-school suspension) for students in schools that participated in Pilot Test II.

We used linear regression models to explore the extent to which scales on the student survey predicted student PSAT scores, the probability of being suspended, and the extent to which student responses differed based on race/ethnicity. We also investigated the extent to which student responses differed based on school affiliation. We did not adjust for multiple comparisons, given the exploratory nature of these analyses. In all regressions, we incorporated school fixed effects to absorb the effects that were particular to each school.

For the teacher and school leader survey samples (Pilot Test II, ATP survey, and New England school leader survey respondents), we used linear regression models to investigate the extent to which scale scores differed based on school characteristics (i.e., Title I status and racial/ethnic composition of the student body) as suggested by the aforementioned research that lower-income and minority students

in low-achieving schools might have fewer SCL opportunities.

Reliability/precision evidence. We used these same analytic files to examine reliability and precision: the extent to which the scales were precise and relatively free from measurement error. For the student survey, for which we had multiple respondents from the same schools, we calculated two types of reliability. First, we examined the extent to which each student responded similarly across items in a given scale. This is called *internal consistency*, and we measured this using Cronbach's alpha. Cronbach's alpha coefficients range between 0 and 1, with values closer to 1 indicating higher reliability and lower measurement error. Researchers often interpret alpha coefficients greater than 0.70 to suggest that survey scores provide relatively consistent and precise scores (Nunnally, 1978). However, this threshold is context dependent, and determinations about whether scores are sufficiently precise depend ultimately on how scores will be used and the associated consequences (Kane, 2011).

Second, we calculated a measure of inter-rater reliability for students within the same school. Inter-rater reliability describes the extent to which two or more students who are observing the same school-based phenomenon agree in their ratings. We measured inter-rater reliability using a coefficient that is known in the literature as *awg*. Like Cronbach's alpha, *awg* ranges between 0 and 1, with values closer to 1 indicating stronger agreement among students. We estimated *awg* for all of the survey items. We hypothesized that *awg* values should be systematically higher for scales measuring school climate and context than for scales that are designed to measure individual student experiences. For the instructional staff and school leader surveys, we did not have sufficient data to estimate inter-rater reliability—therefore, we calculated only internal consistency for those scales.

Usability evidence. Finally, to assess whether the MISCL Toolkit was used as intended, we analyzed qualitative data collected through our observations of students and school staff using the User Guide, Support Tools, and Walkthrough Guide, along with cognitive interviews about their experiences examining the collected data and using the Toolkit's Support Tools and User Guide. The results of this analysis were used to inform revisions to make the Toolkit more user-friendly and useful for school staff. Although the Walkthrough Guide is an instrument for measuring the extent of SCL, we collected only limited information about the validity of the Walkthrough Guide through cognitive interviews and thus focused on usability findings for this instrument.

Results

This section describes the results of our analyses. We begin by describing our efforts to assess validity of survey content through our literature review and the feedback we received from experts. We then describe the validity, reliability, and usability evidence we collected for each element of the Toolkit. Our findings provide some insight about whether the instruments measured what they were intended to measure: the extent of SCL in high schools. In addition, they provide information about whether the MISCL Toolkit was used as intended—to assess, understand, and reflect on the extent of SCL opportunities in classrooms.

Validity Evidence Related to MISCL Toolkit Content

How Did We Collect Evidence on Content Validity?

We collected two types of evidence to assess validity of survey content in the tool development phase: a literature review and feedback from our Advisory Board. In this section, we describe the results of those activities and how they influenced development of the MISCL.

What Did We Learn?

Literature Review

Prior to developing the MISCL surveys, we conducted a literature review to determine the appropriate content to include.[6] The purpose of the literature review was to (1) establish the SCL strategies and contextual conditions on which the survey would focus and (2) review the relevant research on those strategies and conditions, documenting any relationships between

Overall, we found that the effectiveness of SCL in increasing student engagement and academic achievement is mixed.

these strategies and conditions and positive student outcomes (i.e., improved academic performance and student engagement). The literature review was guided by a conceptual framework for measuring SCL informed by NMEF's priorities and experience working with school systems to implement SCL and our experience studying aspects of SCL across many schools. Specifically, the literature review focused on evidence of implementation and outcomes related to comprehensive SCL interventions and those related to our conceptual framework, which focused on SCL as it relates to personalization, competency-based learning, anytime/anywhere learning, data use, and student agency.

We reviewed 225 studies, 156 of which were ultimately included in our review because they provided evidence regarding successes, challenges, and outcomes related to SCL interventions in schools.[7] Our review suggested that the contextual conditions for SCL programs are likely to play an important role in successful implementation, given that most SCL strategies involve a considerable shift from traditional approaches to schooling. In particular, the review pointed to the importance of following three contextual conditions—(1) systems for continuous improvement; (2) people, policies, and infrastructure that emphasize and support SCL; and (3) warm, caring learning environments—that enable SCL to take place in schools.

We examined the relationship between SCL strategies and student outcomes to better understand the extent to which validity evidence on the relationship between SCL and external variables might already exist (and thus make sense to measure as part of the MISCL Toolkit). Overall, we found that the effectiveness of SCL in increasing student engagement and academic achievement is mixed. Some studies found evidence that large-scale SCL interventions can be effective but noted that SCL can be implemented in a wide variety of ways, and not all SCL programs include the five key strategies discussed earlier. We did find some evidence of specific SCL practices that can improve students' academic achievement and engagement, including the following:

- personalization based on student interests (Walkington, 2013; Ku and Sullivan, 2002; Awofala, 2016)
- some competency- and mastery-based systems in which students have unlimited time to master specific learning targets (Kulik, Kulik, and Cohen, 1979; Kulik, Kulik, and Bangert-Drowns, 1990; Abakpa and Iji, 2011; Friedlaender et al., 2014)
- programs linking community service to classroom instruction (Furco, 1996; Billig, 2000; Billig, 2009; Conway, Amel, and Gerwien, 2009; Celio, Durlak and Dymnicki, 2011)
- approaches for teaching metacognitive strategies that help students plan and monitor their own learning (Cross and Paris, 1988; Cardelle-Elawar, 1992; Boulware-Gooden et al., 2007).

That said, the literature is clear that there is high variability in implementation of SCL practices, which may be responsible for the mixed results in some settings. Thus, specific implementation features, choices, and context might matter a great deal for the success of particular SCL approaches.

Advisory Board Feedback

Our seven-member Advisory Board included practitioners (e.g., teachers, principals, technical support providers) and researchers with expertise in studying, measuring, or implementing SCL. We also sought feedback from NMEF staff. We convened two meetings of the full Advisory Board as we were developing the conceptual framework and drafting survey instruments to gather information on strategy wording and contextual conditions we identified

and our draft survey instruments. We also sought just-in-time feedback from individual Advisory Board members on elements of the MISCL Toolkit under development, including the User Guide, Walkthrough Guide, and student focus group.

In the course of their review, the Advisory Board members confirmed that they viewed the five strategies and contextual conditions identified in our literature review as part of SCL. They further confirmed that the five strategies and contextual conditions were well represented in the survey instruments and that the questions were worded using language appropriate to the respondent group (e.g., instructional staff, students, school leaders). Key takeaways from their feedback that we incorporated into our framework and instrument revisions are as follows:

- Balance the need to capture adequate detail with burden on tool users, and use clear, concise language for SCL strategies, contextual conditions, and practices that would be understandable to educators not immersed in SCL
- Consider how the Support Tools can measure equity of student learning opportunities and outcomes in an SCL context
- Consider incorporating families and external organizations or partners in student and adult learning
- Gather varied data on SCL from multiple stakeholders within the school system, including students, and take into account nonsurvey sources of evidence, such as classroom walkthroughs
- Ensure that surveys use reference points that are most appropriate for each participant (e.g., that students are asked about the classes they are taking and instructional staff are asked about the classes they teach), taking into account that SCL might vary widely depending on the reference point used (e.g., if the first class of the day is the reference point on which instructional staff are asked to focus, they might give very different responses than if they were asked to respond in regard to all of the classes they teach)
- Include guidance and materials that support practitioners in the use of MISCL Toolkit data to have informed conversations about continuous improvement of SCL in their schools, and ensure that these materials prompt discussions about equity of student learning opportunities.

How Did We Develop and Revise the MISCL Toolkit in Response to What We Learned?

As a result of feedback from our Advisory Board and our literature review, we developed and refined the MISCL instruments to measure five SCL strategies (learning [1] is personalized, [2] is competency-based, [3] occurs anytime, anywhere, [4] promotes student agency and ownership, and [5] is informed by data) and three contextual conditions to support SCL (systems for continuous improvement; people, policies, and infrastructure; positive learning environments). We also developed a User Guide and other resources to support users and help them reflect on the distribution and equity of SCL supports.

Validity, Reliability, and Usability Evidence for MISCL Toolkit

In this section, we present detailed results for each element of the Toolkit: the User Guide and two Support Tools[8] (i.e., Quick-Start Guide, Reflection Conversation Guide) and the six instruments (i.e., student, instructional staff, district leader, and school leader surveys, the Walkthrough Guide, and the student focus group). First, we discuss our usability evidence regarding the User Guide, Support Tools, and Walkthrough Guide. We then present validity and reliability evidence for the student focus group and each survey instrument that was part of the MISCL Toolkit.

MISCL Toolkit: Research Question and Data Collected

Usability: To what extent was the Toolkit usable and useful to school systems that undertook the Toolkit process?

Data collected and analyzed:

- Qualitative analysis of nine interviews with instructional staff and four interviews with school administrators in Pilot Test II schools
- Observations of Toolkit use in Pilot Test II schools

MISCL Toolkit

How Did We Collect Usability Evidence for the Toolkit?

Through Pilot Test II, we gathered evidence of usability for the whole MISCL Toolkit, including administration of the instruments and the clarity of the instructions in the User Guide, Support Tools, and Walkthrough Guide. We specifically interviewed school administrators and instructional staff who facilitated the Toolkit process or had been invited to participate in the data reflection conversations to understand their experiences with the Toolkit and whether they found the Toolkit usable and useful.

What Did We Learn?

We analyzed the interview data pertaining to whether MISCL Toolkit users understood and interpreted Toolkit materials as intended and also whether the materials enabled users to (a) collect systematic information about SCL from a variety of stakeholders within the school and (b) help schools collectively analyze and reflect on that information for continuous improvement of student-centered practices. Although we only tested the Toolkit in two school systems, and findings from this pilot test might not be generalizable to all school systems, we distilled five main lessons from this analysis based on responses of multiple interviewees in both school systems.

First, the school administrators who led the Toolkit administration in Pilot Test II schools used the materials as intended; interpreted the User Guide guidance and suggested steps for goal-setting, data collection, analysis, and reflection consistently; and found the process to be valuable and appreciated the information gathered with the Toolkit. Although school administrators used the Toolkit as intended and found the process valuable, they also reported that the process was complex and time-consuming. School administrators specifically reported that the User Guide was lengthy, and some reported that they did not read all of the materials or use the planning worksheets. In addition, some of the school leaders who led the administration of the Toolkit did not share any written guidance from the Support Tools or User Guide with reflection conversation participants. School administrators reported that they were reluctant to share the Support Tools—particularly the Walkthrough Guide—with instructional staff who were participating in the activity, because they thought it would be too burdensome for instructional staff to take the time to review before participating in the conversations.

Second, instructional staff who participated in the classroom walkthrough reported that they would have preferred clearer communication about the purpose of the walkthrough (i.e., to get a snapshot of SCL practices in different classrooms)—in particular, participants' roles and responsibilities and the walkthrough schedule. For example, some instructional staff whose classrooms were observed reported that they did not know there would be multiple groups of observers rotating through their classrooms. These instructional staff also wondered whether they would receive any feedback from the observations and how it would be delivered. Some instructional staff who observed classrooms reported that they did not know what would happen with the observation notes they compiled and similarly wondered whether the instructional staff of the observed classrooms would receive any feedback. All of that said, participants greatly appreciated the set of 20 example look-fors of SCL strategies that were provided as part of the Walkthrough Guide and thought it was relatively easy to reflect on the presence of those look-fors in the classrooms they observed.

Third, we had originally included an "instructional log" for instructional staff whose classrooms

were observed in the walkthrough. The instructional logs were intended to be an additional measure that would allow instructional staff to report on their use of SCL strategies every day for a period of three days through a five-minute survey. Although the instructional logs were intended to help instructional staff and school leaders gather a comprehensive picture of SCL, the logs themselves took a fair amount of time, given that they had to be administered every day. Furthermore, in the two schools we observed, instructional staff who completed the logs were not asked to share their logs or participate in the reflection conversations about the walkthrough, although the User Guide did direct facilitators to include them if possible. The User Guide might not have been clear enough about the need to include instructional staff whose classrooms were observed in the reflection conversations. In addition, facilitators might have been concerned that instructional staff might feel put on the spot if they were included in a conversation with other participants about their classroom instruction. Although the few instructional staff who completed the logs found them useful for reflection about their own instruction, we also surmised that log data would have been difficult to use and interpret, given that the same measures were repeated—and thus would have to be examined—across several days. Therefore, we decided to remove the logs from the MISCL Toolkit.

Fourth, a majority of MISCL Toolkit facilitators and participants—administrators and instructional staff alike—reported that they would like additional guidance for organizing the reflection conversation and using the Support Tools. In our observations of Toolkit use, we did not see any evidence that the planning materials intended to facilitate the reflection conversation discussion had been used. Interviewees in both Pilot Test II schools confirmed that they did not use these materials, generally because, as they told us, they did not notice them within the User Guide. As a reminder, the Toolkit is intended to be a self-serve set of instruments and guidance that school systems can just use without additional advice or support. Thus, school systems participating in the pilot received very limited direction on how to use the Toolkit in their contexts, so that we could observe the process as it might be used in any school that accesses the Toolkit

Although school administrators used the Toolkit as intended and found the process valuable, they also reported that the process was complex and time-consuming.

in the future. Our data specifically pointed to the following areas where more guidance within the Toolkit materials could be useful:

- **Comparing data.** The interviews and our observations indicated that participants would have liked to compare data across surveys but were not sure how to do so efficiently. In particular, although several items across surveys were similar, they were not necessarily directly comparable given slight differences in response scales. Furthermore, although the User Guide aimed to emphasize the importance of looking at the student survey results to see whether there were differences across student groups (e.g., students in different grade levels or of different races/ethnicities), participants in our two schools did not examine the data in this way.
- **Incorporating student voices.** Students were included in the reflection conversations in both schools, and administrators reported that students' perspectives were important for gaining a complete picture of SCL in the schools. However, in one school, the few students who participated in the reflection conversations generally did not feel comfortable speaking up without being asked, even though including students' perspectives was intended to be a key aspect of the discussion.

Fifth, our analysis of the interviews and observations of Toolkit use suggested the need for the Toolkit Administration Team—including anyone coordinating the use of the Toolkit and facilitating any conversations or aspects of the Toolkit process—to strive as much as possible to set a positive, low-stakes tone for their staff and students when explaining the purpose of the Toolkit. For example, many instructional staff reported that the Toolkit activities—in particular, the survey and the walkthrough—felt evaluative, and they worried about whether their instruction and survey responses would be consistent with administrators' expectations. Some students who participated in the focus group and reflection conversation similarly felt that their comments needed to be consistent with administrators' expectations about the SCL in their schools rather than their actual experiences.

How Did We Revise the MISCL Toolkit in Response to What We Learned?

We made the following key revisions to the User Guide, Support Tools, and Walkthrough Guide based on our analysis of response processes:

- We streamlined the instructions and eliminated redundancies to reduce the length and burden of the User Guide and reflection materials. We also developed a Quick-Start Guide to offer users an overview of the Toolkit.
- We removed the instructional staff logs because they did not seem broadly useful for informing a schoolwide understanding of SCL implementation and were burdensome for instructional staff to complete.
- We amended our Walkthrough Guide to address instructional staffs' desire for clear communication about the purpose, roles, and responsibilities for the walkthrough. This change was made to encourage the lead facilitator to interact with all participants—including those whose classrooms were being observed—prior to the walkthrough to discuss the walkthrough process and how feedback will be shared and to share a schedule and plan for the walkthrough in advance. In addition, the Walkthrough Guide had previously been characterized as a Support Tool instead of an instrument. To incorporate observations from the walkthrough into the reflection conversation, we recharacterized the Walkthrough Guide as an MISCL instrument that should be completed prior to the reflection conversation and discussed alongside findings from the other instruments.
- We put together a Reflection Conversation Guide that is a more-integrated part of the Toolkit to address the need for more guidance about how to organize and facilitate the reflection conversation and support participants to compare data. (Previous guidelines were in an appendix to the Toolkit and not explicitly emphasized as a Reflection Conversation

"I wanted a voice for the students; it's not the same hearing from the teachers, as hearing from the students. [Students] have a different perspective of what we believe we're doing, so that's what I wanted to show the leadership: the student voice."

—School administrator

Guide.) Also, we added some additional information to the Quick-Start Guide and User Guide on how facilitators might best lead those conversations.
- We provided guidance on setting a positive, nonevaluative tone in the Quick-Start Guide. We also added tips to the User Guide about transparency, tone, and communication.
- We reviewed the instructional staff and student surveys side by side and made small line edits for consistency in wording to address the difficulty that MISCL Toolkit users had with comparing student and instructional staff survey responses. In addition, given that each section of each of the surveys was intended to measure a particular SCL strategy, we added a question to the end of each section of the student survey and the instructional staff survey asking respondents to provide a holistic rating of their agreement about the extent to which their classes reflected that SCL strategy. For example, at the end of the section on personalization, we asked students to indicate their agreement with the statement, "Instruction in my classes is personalized," taking into account the responses they had given on questions in that section of the survey. Similarly, at the end of the personalization section of the instructional staff survey, we asked instructional staff about their agreement with the statement, "My instruction is personalized," given how personalization was measured in that section of the survey. We added a similar item to the end of some sections of the school and district survey to allow for more comparisons among items from different surveys.

Student Focus Group

How Did We Collect Validity Evidence for the Student Focus Group?

The student focus group consisted of a series of questions about the SCL strategies that were part of our framework. The questions addressed, for example, whether and how instructional staff address students' individual learning needs, and the extent of student decisionmaking in the school. These questions were intended to provide rich examples of students' SCL experiences to supplement survey data. Following an observation of the student focus group in action with a school facilitator, we conducted another focus group with the students—and an interview with the facilitator—to gather information about whether the focus group questions were easy to understand and interpreted as intended.

What Did We Learn?

We came away with two main lessons from this analysis and identified some misunderstandings about particular focus group questions, which we addressed with minor edits to the focus group protocol. First, students and facilitators in both schools understood most of the questions and reported that the language was relevant to their school contexts. Facilitators found the instructions for facilitating the group to be helpful and thought they worked well, and students agreed. Students felt comfortable answering the questions honestly and thought the discussion, which took about an hour, was an appropriate length.

Second, the need for MISCL Toolkit facilitators to clearly communicate the purpose of the student focus group to students was highlighted by the differences in how students in each of the two schools perceived the information would be used. At one school,

Student Focus Group: Research Question and Data Collected

Response processes: To what extent did the students and the MISCL facilitator find the student focus group questions easy to understand and interpret them as intended?

Data collected and analyzed:

- Qualitative analysis of two focus groups with ten students who participated in the focus groups and two interviews with the adults who facilitated the groups across the two Pilot Test II schools
- Observations of student focus groups as they were conducted in the two Pilot Test II schools

students reported that the facilitator was trying to use the focus group to better understand students' experiences; at the other school, students felt that the facilitator was using the group to explain the school's SCL model to the RAND research team. These differences in students' perceptions of the purpose of the focus group might have had an impact on how students responded to question prompts. Specifically, if students feel that the focus group is less about understanding students' experiences and more about confirming a school's SCL approach, they might not give candid responses to focus group questions.

Nonetheless, students in both schools emphasized their belief that their perspectives are important, because they felt that their experiences of SCL might be very different from those of instructional staff and administrators at their schools. Instructional staff and administrators in both schools agreed that incorporating students' perspectives in the MISCL Toolkit was important.

How Did We Revise the Student Focus Group in Response to What We Learned?

We revised the student focus group protocol slightly to correct a few misunderstandings that were raised by students in the focus group. In addition, the Toolkit now includes some additional guidance to users about how to communicate the purpose of the student focus group, and we underscore that facilitators should be clear to students about the desire to learn from their honest responses to focus group questions. Because students might not participate in the reflection conversation, student feedback from the focus group could help explain differences in students' and instructors' survey responses. In the Reflection Conversation Guide, we now instruct the lead facilitators to distribute notes from the student focus group to reflection conversation participants and to incorporate student comments into discussion of survey results.

Student Survey

How Did We Collect Validity Evidence for the Student Survey?

Beyond our process to collect expert feedback, we gathered validity evidence related to students' response processes through focus groups with students who had completed the survey. We also examined evidence on the internal structure and reliability of the survey, and any relationships between what we measured in the student survey and other variables, through quantitative analyses of the survey responses from students as part of our Pilot Test II.

What Did We Learn?

Response processes. Using focus group data from students who completed the survey, we came away with two main takeaways, which we addressed through some revisions to survey content. First, students stated that they understood the questions, that the language was relevant to their school contexts, and that response scales were understandable and appropriate across the two Pilot Test II schools. Students' reported interpretation of most questions was consistent with the intended interpretation, and points of confusion were easily resolved with minor edits to selected survey items. However, several students said they would have answered the questions differently for their elective classes than they would for their academic classes. Second, most students opined that the survey was too long, and that several of the questions were redundant.

Internal structure. The student survey contained a total of 27 scales: five for personalization, three for competency-based learning, four for anytime/anywhere learning, seven for student agency and ownership, two for learning informed by data, and six measuring school contextual conditions. Complete lists of scale names, descriptive statistics, and items can be found in Tables B.1 through B.4 in Appendix B.

Because the items in a particular scale are all assumed to be associated with the same construct, we expected that responses to these items would have similar patterns. For example, the student survey was intended to measure multiple key constructs within the categories of personalization, competency-based

learning, anytime/anywhere learning, student agency and ownership, and data use. We expected students to respond similarly to the items about personalization because the items probed about aspects of the same construct. We used CFA to test whether there was empirical evidence that supported the hypothesis that items intended to measure the same construct were related in anticipated ways.

Based on the diagnostic information described earlier in this report, CFAs suggested that some scales had more integrity than others. Of the 27 scales, the following details emerged:

- Seven provided strong or adequate evidence that students responded similarly to items and/or that the items were closely related.
- Seven scales contained three or fewer items and were thus untestable.
- Thirteen yielded empirical evidence that raised questions about whether the items were closely related and measuring a common construct. In these cases, following the procedures described earlier in this report, we used the CFA results diagnostically to revise seven of these scales, which improved their integrity. Of the 13 scales, there were five scales that we did not revise, despite the fact that some information suggested that the scales were not functioning as intended. However, the diagnostic information about these scales was mixed, with some information suggesting adequate fit (CFI and SRMR) and other information suggesting poor fit (RMSEA). Some research suggests that RMSEA may sometimes be inflated if sample sizes are small (Kenny, Kaniskan, and McCoach, 2015), and so we interpret the evidence holistically to suggest that there is adequate evidence supporting these scales.

Future survey administrations will offer additional opportunities to explore the integrity of these scales further. The CFA results for the revised scales are presented Tables B.1 through B.4 in Appendix B.

Following these CFAs, we computed scale scores for each of the 27 scales and then estimated the correlations among these scale scores to inspect the extent to which the scales were related in anticipated ways. Scales that measured aspects of schools that were hypothesized to support SCL practices were positively associated, and the correlations tended to be moderate in magnitude. Most scales correlated between 0.19 and 0.51, which provides evidence that the scales are measuring distinct but interrelated aspects of SCL practices and school culture. Correlations are reported in Table B.12 in Appendix B.

Relationships with external variables. We used linear regressions to test the a priori hypotheses that SCL practices would be (1) positively associated with student achievement; (2) negatively associated with discipline incidents; and (3) negatively associated with student characteristics; and that (4) the school with more SCL experience would have comparatively higher scale scores and (5) students in higher grades

Student Survey: Research Questions and Data Collected

Response processes: To what extent did students find the survey questions easy to understand and interpret them as intended?

- Qualitative analysis of two student focus groups, with eight students total in the two Pilot Test II schools

Internal structure: To what extent did responses to survey items measuring the same topics relate? To what extent did scales correlate in ways that were consistent with theory?

- Factor analyses of students' responses
- Correlations of scale scores

Relationships with external variables: To what extent were students' responses related to external variables in ways that were consistent with theory?

- Relationships with demographic and achievement variables

Reliability: To what extent were items and scale scores consistent and free of measurement error?

- Internal consistency
- Inter-rater reliability

Our analysis of student suspensions found that students who reported more student-centered practices and contextual supports were generally less likely to be suspended, a result that was consistent with our expectations.

would report greater engagement with SCL practices. We conducted these linear regressions with each of the 27 survey scales. We used PSAT scores as our measure of student achievement and student suspensions as our measure of discipline.

In our analysis of PSAT scores, we found that, in general, most results were not statistically significant, but the direction of the coefficients suggests that students who reported more student-centered practices and supports had lower PSAT scores. This was not consistent with our hypothesis. It is important to note that the PSAT has many limitations as a measure of student achievement and might not be sensitive to student learning or to SCL-based practices, and, therefore, these findings should be interpreted with caution. Future data collection should include other measures of student achievement to provide richer evidence to support or reject the research-based hypothesis that SCL practices should be associated with higher student achievement.

Our analysis of student suspensions found that students who reported more student-centered practices and contextual supports were generally less likely to be suspended, a result that was consistent with our expectations. In general, the regression coefficients were negative, and nearly half of these coefficients were statistically significant at the .05 level. However, there were very few students who responded to the survey who had also been suspended; this is consistent with the small size of both schools and the schools' likely adoption of progressive discipline policies that limit use of suspension.

In general, we found no significant differences in students' responses to survey items in regard to SCL based on race/ethnicity or FRL eligibility. There were largely no significant differences based on grade level, although students in grade 12 did perceive that they had fewer opportunities to engage in mastery-based learning or to receive course credit based on mastery, a finding that was also not consistent with our expectations.

In terms of comparisons between the two schools, we did find significant differences, consistent with our expectations. In general, students at the school that school leaders described as an experienced SCL school had higher ratings of their school's SCL practices, compared with the school that school leaders stated was newer to SCL. Full regression results are presented in Tables B.8 and B.9 in Appendix B.

In sum, although the SCL surveys likely reflect something about the extent of SCL in schools, as we have defined it, the relationships between SCL and external variables did not always align with our hypotheses or expectations, which raises questions about whether SCL should always be expected to support improvements to student achievement or may be meeting other goals beyond those associated with achievement. On the other hand, our findings are limited by the small survey samples in just two school contexts. These results are considered further in the conclusion to this report.

Reliability. As described earlier in this report, we used two measures of reliability for the student survey. Cronbach's alpha, our measure of internal consistency, measures the extent to which each student responded similarly across items in a given scale. The alpha reliability coefficients were above 0.70 for all 27 scales on the student survey, suggesting that there was strong internal consistency of items within the scales. Table

B.1 in Appendix B presents alpha reliability coefficients for each of the scales.

In terms of inter-rater reliability, we found that, consistent with our expectations, students had more-strongly shared perceptions of school-level phenomena than phenomena that describe individual experiences. For example, the average *awg* for the items in the contextual conditions supporting SCL area was the highest of the six measured areas on the student survey (*awg* = 0.58), suggesting relatively strong agreement and shared perceptions. This is consistent with the fact that the contextual conditions supporting SCL items focus on school-level or collective phenomena (e.g., "All students are encouraged to go to college.") On the other hand, the average *awg* for the Data Use items was 0.17, suggesting that students tended not to agree with one another on these items. This is unsurprising, given that many of the items in the Data Use area focus on individual student experiences and actions (e.g., "I check on my progress at least a few times a year"). Detailed inter-rater reliability results can be found in Appendix Table B.1.

How Did We Revise the Student Survey in Response to What We Learned?

Using our analysis of response processes, we made the following key revisions to the student survey:

- We revised the survey to clarify that students should keep their academic classes in mind when responding to survey questions to address the comment that students would have answered the questions differently for their elective classes than they would for their academic classes.
- We reviewed the survey closely for redundancies and omitted items that overlapped or were similar to others to address students' concern about the length of the survey and the similarities between some items. We also used the factor analyses (see discussion of internal structure and reliability for more detail) to supplement our review for content and identify redundant items for deletion.
- Diagnostic evidence from the factor analyses suggested that some pairs of items were correlated more strongly than could be explained by the fact that the items measured a single construct. In these cases, we reviewed the survey items and flagged items that were redundant in content. These items were subsequently removed from the survey instruments. In other cases, it was determined that although subsets of items overlapped considerably, they still captured something important about the construct. In these cases, we left the items in the survey, but acknowledged the excess item overlap by allowing specific pairs of items to correlate with one another.

> We found no significant differences in students' responses to survey items in regard to SCL based on race/ethnicity or FRL eligibility.

As a final step, we reviewed the student and instructional staff surveys for alignment and consistency in item content and wording and made further minor revisions as needed.

Instructional Staff Survey

How Did We Collect Validity Evidence for the Instructional Staff Survey?

We gathered validity evidence related to teachers' response processes through interviews with instructional staff who completed the survey. We also examined evidence on internal structure and reliability of the instructional staff survey, and any relationships between survey measures and external variables, through various quantitative analyses of the survey responses from teachers who participated in the second pilot, along with responses from a fielding of the instructional staff survey to the RAND ATP.

Instructional Staff Survey: Research Questions and Data Collected

Response processes: To what extent did instructional staff find the survey questions easy to understand and interpret them as intended?

- Qualitative analysis of 13 instructional staff interviews

Internal structure: To what extent did responses to survey items measuring the same topics relate? To what extent did scales correlate in ways that were consistent with theory?

- Factor analyses of responses of instructional staff from Pilot Test II schools and ATP
- Correlations of scale scores

Relationships with external variables: To what extent were instructional staff responses related to other external variables in ways that were consistent with theory?

- Relationships with school characteristics for Pilot Test II instructional staff and ATP teachers

Reliability: To what extent were items and scale scores consistent and free of measurement error?

- Internal consistency of responses of instructional staff from Pilot Test II schools and ATP

What Did We Learn?

Response processes. Using interview data from teachers who completed the instructional staff survey, we derived three main takeaways, which we addressed through revisions to the survey. First, several instructional staff reported feeling intimidated by the survey content. The surveys included several SCL practices that some instructional staff had never heard of or never engaged in. In the cognitive interviews, instructional staff questioned the need to ask about those practices and stated that they felt uncomfortable responding to those items because they felt they could not give an honest response about the extent to which they had engaged in those practices. Instructional staff may have been concerned about judgment or evaluation from their leaders. Instructional staff responses to some survey items could also be driven by the social desirability of more-positive answers, which could potentially inflate reports of SCL.

Second, some of the more-complex survey items intended to help instructional staff and school leaders gather a comprehensive picture of all aspects of instructional staff perceptions and practices related to SCL did not appear to be useful to school staff. The more-complex survey items asked instructional staff to rate the extent of SCL in a series of hypothetical vignettes and then rate their own instruction, and took some time to answer. Furthermore, neither of the schools that tested the Toolkit used the data from the vignette survey questions in their reflection conversations. We surmised that the complexity of analyzing and interpreting these items was the main barrier to school staff using them for analysis and that the value added by the precision of these items was not justified because they proved to be so difficult for school staff to use.

Third, some instructional staff survey questions referenced the "first academic class of the day." Many instructional staff had trouble answering those questions because their first academic class was reportedly atypical of their regular instruction.

Internal structure. The instructional staff survey contained a total of 30 scales: four for personalization, three for competency-based learning, three for anytime/anywhere learning, five for student agency and ownership, five for learning informed by data, and ten measuring school contextual conditions. Complete lists of scale names, descriptive statistics, and items can be found in Table B.2 in Appendix B.

As with the student survey, we conducted CFA separately for each of these scales to find evidence that instructional staff responded similarly to items designed to measure a common construct. Of the 30 scales:

- Ten provided strong or adequate evidence that instructional staff responded similarly to items or that the items were closely related.
- Seven scales contained three or fewer items and thus were not testable.
- Thirteen provided some empirical evidence that raised questions about whether the items were closely related and measured a common construct. In these cases, following the procedures described earlier in this report, we used the CFA results diagnostically to revise the scales. Again, we reviewed the survey items and flagged items that were redundant in content. These items were subsequently removed from the survey instruments. Four survey scales were subsequently revised, either by revising the CFA models or by removing redundant items, which improved the integrity of the scales. Detailed CFA results for each of the revised scales can be found in Table B.6 in Appendix B. There were ten scales that we did not revise, despite the fact that some information suggested the scales were not functioning as intended. However, as with the student survey, the diagnostic information about these scales was mixed, with some information suggesting adequate fit (CFI and SRMR) and other information suggesting poor fit (RMSEA). Some research suggests that RMSEA may sometimes be inflated if sample sizes are small (Kenny, Kaniskan, and McCoach, 2015), and so we interpreted the evidence holistically to suggest that there is adequate evidence supporting these scales.

Future survey administrations will offer additional opportunities to explore the integrity of these scales. Detailed CFA results for each of the revised scales can be found in Table B.6 in Appendix B.

Following these CFAs, we computed scale scores for each of the 30 scales and then estimated the correlations among these scale scores to inspect the extent to which the scales were related in anticipated ways. Consistent with our expectations, scales that measured aspects of schools that were hypothesized to support SCL practices were moderately and positively associated; the average correlation was approximately 0.24, which provided evidence that the scales measured distinct but related aspects of SCL. Also consistent with our expectations, the three scales that addressed the inadequacy of resources showed systematically negative correlations with other scales in the survey.

> Several instructional staff reported feeling intimidated by the survey content. . . . Instructional staff may have been concerned about judgment or evaluation from their leaders.

Relationships with external variables. We used linear regressions to test the a priori hypotheses that SCL practices would be (1) negatively associated with the proportion of students from racial minorities and (2) negatively associated with socioeconomic status (lower–socioeconomic status schools would have less use of SCL practices) and that (3) the school with more SCL experience would have comparatively higher scale scores. We conducted these linear regressions with each of the 31 survey scales. We used the proportion of black and Hispanic students as our measure of school racial composition and Title I status as our measure of socioeconomic status.

For the instructional staff survey scales, we did not explore associations with student outcomes, because the majority of instructional staff in these analyses came from the ATP. This created an analytic issue, in that we did not have access to achievement data for ATP teachers. In addition, although we did have achievement data for the subset of teachers who participated in our pilots, those data were only

In general, instructional staff at the school that had described itself as an SCL school reported higher ratings of their school's SCL practices compared with the school that had described itself as newer to SCL.

available at the school level. Given that high school students take courses from multiple instructional staff, and that instructional staff were prompted to answer survey questions with a specific class in mind, we determined that it was not appropriate to make inferences regarding the relationship between instructional staff survey reports and student achievement based on school-level student achievement data.

Using data collected from our two Pilot Test II schools and the ATP, we found only a very small number of statistically significant differences in responses based on Title I eligibility, some of which were positive (Title I–eligible schools had higher average survey responses) and some of which were negative (Title I–eligible schools had lower average survey responses). Taken together, the regression results suggest that there were no discernible differences based on Title I eligibility. One possible explanation for this finding is that Title I status, which is a schoolwide indicator, is only weakly related to the experiences of one specific instructional staff member. Finer-grained information about socioeconomic status, or, alternatively, data from a representative sample of teachers within each school, may shed important additional light on these relationships.

In terms of racial composition of the student body, we also found few statistically significant differences. Noticeably, all of the scales about the inadequacy of resources to support the implementation of SCL practices showed positive associations with the proportion of students in racial minority groups, meaning that schools with higher shares of minority students typically reported having fewer resources to support SCL. This is consistent with our a priori hypotheses and prior research.

Finally, in terms of comparisons between the two Pilot Test II schools, we did find significant differences, consistent with our expectations and with the student survey results. In general, instructional staff at the school that had described itself as an SCL school reported higher ratings of their school's SCL practices, compared with the school that had described itself as newer to SCL. Full regression results are presented in Table B.10 in Appendix B.

Reliability. As described earlier in this report, we calculated the internal consistency of each of the scales in the instructional staff survey. Overall, the calculated Cronbach alphas for those scales ranged from 0.42 to 0.93, and all but three of the scales had estimated reliability coefficients that were above 0.70. Notably, two of the scales with lower reliability had three or fewer items. Internal consistency is a function of the number of items in a scale (Brennan, 2001), so this is not unexpected. One scale with six items had lower internal consistency (alpha = 0.60); however, on further inspection, this seemed to be driven largely by the fact that there was little variance across teachers rather than a larger-than-anticipated amount of measurement error. Taken together, among these results, there was strong internal consistency of items within the scales. Detailed results can be found in Table B.6 in Appendix B.

How Did We Revise the Instructional Staff Survey in Response to What We Learned?

Using our analysis of response processes, we made the following revisions to the instructional staff survey:

- To make the survey feel less intimidating and evaluative to instructional staff, we revised the introductory survey language to stress that staff should not feel that they need to be familiar with all of the practices in the survey and that the survey is not intended to be evaluative. We also revised response choices with the word "never," replacing it with "not yet" or "not at all," in case instructional staff felt self-conscious about indicating that they "never" engaged in certain SCL activities.
- We removed the vignettes from the surveys because they were too complex for school staff to analyze.
- We asked instructional staff to respond about a typical class they teach to address their concerns that the first academic class of the day varied and was not always representative of most of the classes they teach. In the revised instructional staff survey, a *typical class* is defined as the class that an instructional staff member believes best exemplifies the instruction they provide most often or for the most students they teach.
- Diagnostic evidence from the factor analyses suggested that some pairs of items were correlated more strongly than could be explained by the fact that the items measured a single construct. In these cases, we reviewed the survey items and flagged items that were redundant in content. These items were subsequently removed from the survey instruments. In other cases, it was determined that although subsets of items overlapped considerably, they still captured something important about the construct. In these cases, we left the items in the survey, but acknowledged the excess item overlap by allowing specific pairs of items to correlate with one another.

Schools with higher shares of minority students typically reported having fewer resources to support SCL.

District and School Leader Surveys

How Did We Collect Validity Evidence for the District and School Leader Surveys?

We gathered validity evidence related to districts' or school leaders' response processes through interviews with leaders who completed the survey. We also examined evidence about internal structure and reliability of the survey, and any relationships between survey measures and external variables, through various quantitative analyses of the survey responses from leaders who participated in Pilot Test II, and from a sample of New England district or school leaders.

What Did We Learn?

Response processes. Given that we did not have as many interviews with school and district leaders as we did with instructional staff, we discuss the themes that emerged across school and district leader interviews together. We conducted cognitive interviews with leaders who had completed the leader surveys to examine the extent to which they understood and interpreted the survey items as intended. Some misunderstandings about particular survey items were addressed with minor edits to survey content. We found that most school and district leaders felt that the surveys were straightforward and easy to complete. Unlike instructional staff, leaders—on the whole—did not share concerns that the survey felt evaluative or was intimidating in its content, although leaders noted that respondents might not be able to answer all items easily, depending on their progress on and understanding of SCL.

District and School Leader Surveys: Research Questions and Data Collected

Response processes: To what extent did school and district leaders find the survey questions easy to understand and interpret them as intended?

- Qualitative analysis of seven school and district leader interviews

Internal structure: To what extent did responses to survey items on the school leader survey measuring the same topics relate? To what extent did scales correlate in ways that were consistent with theory?

- Factor analyses of responses from New England school leaders
- Correlations of scale scores for New England district and school leaders

Relationships with external variables: To what extent were school leader responses related to other external variables in ways that were consistent with theory?

- Relationships with school characteristics for New England school leaders

Reliability: To what extent were items and scale scores consistent and free of measurement error?

- Internal consistency for New England district and school leaders

Internal structure. We identified 28 scales consisting of multiple items measuring key constructs associated with the five SCL strategies within the school leader survey. There were three personalization scales, four competency-based learning scales, three anytime/anywhere learning scales, one student agency and ownership scale, six learning informed by data scales, and 11 scales measuring school and contextual conditions. We conducted all of our school leader and district leader analyses using data collected from our two Pilot Test II schools, merged together with data from administration of the survey to school and district leaders in the New England region.

Of the 28 scales:

- Seven provided strong evidence that school leaders responded similarly to items and that the items were closely related.
- Four contained three or fewer items and were thus untestable.
- Seventeen provided some empirical evidence that raised questions about whether the items were closely related and measured a common construct. In these cases, following the procedures described earlier in this report, we used the CFA results diagnostically to revise the scales. Again, we reviewed the survey items and flagged items that were redundant. Fifteen survey scales were subsequently revised by revising the CFA models. In some cases, overlapping items were allowed to correlate. Detailed CFA results for each of the revised scales can be found in Appendix B, Table B.7.

Even after model revisions, two scales did not function well and did not show evidence that the items were closely related, leaving 26 scales that had evidence supporting their structure. We do not recommend using the two scales that did not function well; however, the items from these scales may be informative, and so they can be used in single item analyses. Detailed CFA results can be found in Table B.7 in Appendix B.

Following these CFAs, we computed scale scores for each of the 26 scales and then estimated the correlations among these scale scores to inspect the extent to which the scales were related in anticipated ways. Consistent with our expectations, scales that measured aspects of schools that were hypothesized to support SCL practices were positively associated, which provides evidence that the scales are measuring distinct but related aspects of SCL. Also consistent with our expectations, the nonmastery-based learning scale was systematically negatively associated with other scales in the survey. Contrary to our expectations, questions in the school leader survey about inadequate resources to implement SCL were not negatively correlated with scales that described

District leaders' perceptions of SCL practices were strongly negatively associated with their perceptions of the extent to which there were inadequate resources available to support these practices.

SCL practices. Although we cannot be sure why this was the case, it is possible that school leaders, unlike instructional staff and district leaders, did not perceive these conditions as being barriers to SCL implementation.

We identified 20 scales in the district leader survey, including three personalization scales, two anytime/anywhere learning scales, one student agency and ownership scale, four learning informed by data scales, and ten scales measuring school and contextual conditions. Because of the limited sample size, we did not conduct CFA analyses on the district leader surveys; however, we did examine the scale correlations. These correlations showed that scales were related in anticipated ways. In particular, district leaders' perceptions of SCL practices were strongly negatively associated with their perceptions of the extent to which there were inadequate resources available to support these practices. Correlations are reported in Tables B.14 and B.15 in Appendix B.

Reliability. All but five of the 26 scales on the school leader survey had estimated reliability coefficients that were above 0.70, and three of the scales with low reliability had three or fewer items. On the district leader survey, all but two of the estimated reliability coefficients were above 0.70, and the scales with lower reliability had three or fewer items. Taken together, this suggests that there was strong internal consistency of items within the scales. Detailed internal consistency results can be found in Appendix B, Table B.3 and Table B.4.

Relationships with external variables. We used linear regressions to explore the association of the school leader survey scales and school characteristics (including Title I eligibility and the proportion of students that were identified as black or Hispanic). As noted, these analyses were not conducted for the district sample given its small size. We did not explore associations between the school leader survey scales and student outcomes, because most of the participating schools did not have publicly available achievement data. For the school leader survey, we found very few statistically significant regression coefficients and no consistent findings suggesting that, based on reported practices from school leaders, Title I schools or schools serving higher proportions of black and Hispanic students had systematically different SCL practices than other schools. Although this result is not consistent with our a priori hypotheses, it is not possible to determine whether the null findings reflect the fact that there are no relationships, or the fact that our sample size was small and we might not have had adequate power to detect this subgroup's differences. Full regression results are presented in Table B.11 in Appendix B.

How Did We Revise the District and School Leader Surveys in Response to What We Learned?

Using our analysis of response processes, we made the following revisions to the surveys:

- We revised the introductory language in the school and district leader surveys to be consistent with revisions to the instructional staff survey. Those revisions were made to the instructional staff survey so it would feel less intimidating and evaluative. We reasoned that these changes could also make the school and district leader surveys less intimidating and encourage leaders to answer the surveys honestly. Specifically, we revised the introductory survey language to stress that leaders should not feel that they need to be familiar with all

of the practices in the survey and that the survey is not intended to be evaluative.

- We added a question to the end of some sections of the surveys asking respondents to provide a holistic rating of their agreement about the extent to which instruction at their school reflects that aspect of SCL, to make comparisons among the instructional staff, leader, and student surveys easier. For example, at the end of the section on personalization in the school leader survey, we asked school leaders about their agreement with the statement, "Instruction at my school is personalized," given how personalization was measured in that section of the survey. We added a similar item to the end of each applicable section of the district leader survey, and also to the instructional staff and student surveys, as described earlier. In some cases, we had only asked a small number of questions in the school or district leader survey—not enough to cover the full construct being measured in that section. In those instances, we did not include an end-of-section question, given that the respondent would not be able to give a holistic rating of that construct based on how it was measured in a particular section. For example, in the district leader survey, we asked only a few questions on competency-based learning, anytime/anywhere learning, and student agency. Thus, we did not include end-of-section questions on these aspects of SCL.
- Diagnostic evidence from the factor analyses suggested that some pairs of items were correlated more strongly than could be explained by the fact that the items measured a single construct. In these cases, we reviewed the survey items and eliminated one of the redundant items. In other cases, we had evidence that the scales were not functioning as intended. In these cases, we left the items in the survey for individual item analysis.

Conclusions and Implications

This report presents evidence on the validity, reliability, and usability of the MISCL Toolkit, which is intended to enable high schools to assess, understand, and reflect on the extent of SCL and supports for SCL in high schools. The two main research questions explored in this report are as follows:

- To what extent did the MISCL instruments measure what they were intended to measure, drawing on evidence related to content, response processes, internal structure, relationships with external variables, and reliability?
- To what extent was the MISCL Toolkit usable and useful to those who undertook the Toolkit process?

We explored these questions through a variety of activities, including a literature review and collection of expert feedback, pilot tests of the Toolkit instruments in four different school contexts, administration of the survey instruments to supplementary samples of district or school leaders and teachers beyond those in the four school contexts,

Advisory Board feedback helped us revise the MISCL instruments and create a User Guide that describes SCL in clear, understandable terms that would make sense to schools with a variety of SCL experience.

and observations and interviews with those using the MISCL Toolkit in two school contexts.

Summary of Key Findings

To What Extent Did the MISCL Instruments Measure What They Were Intended to Measure?

Our work suggests the following main takeaways regarding the reliability and validity of the MISCL Toolkit instruments:

Evidence suggested MISCL Toolkit instruments measured SCL constructs and school-level SCL. Our literature review found evidence that specific SCL practices measured through our instruments could improve students' academic achievement and engagement in school. That said, research also emphasized that the implementation of SCL interventions can be highly variable, which likely will have an impact on student outcomes.

The members of our Advisory Board, which consisted of practitioners and researchers with expertise on measurement, SCL, and continuous improvement in schools, opined that our instruments appeared to measure the aspects of SCL we intended. Their feedback also helped us revise the MISCL instruments and create a User Guide that describes SCL in clear, understandable terms that would make sense to schools with a variety of SCL experience; emphasizes equity of SCL opportunities for students; and incorporates multiple stakeholder voices into the measurement and improvement process.

Evidence we gathered in regard to response processes—including cognitive interviews and focus groups with Pilot Test II school administrators, instructional staff, and students—indicated that survey respondents understood the items in our surveys and interpreted them in the way that we intended. Sometimes, respondents felt that a question would be easier to answer if we provided them with a different reference point (e.g., all of their courses instead of their first academic class of the day), which helped us revise our items to point to the clearest points of reference.

Evidence based on internal structure suggested that some but not all items in scales were closely related, and instrument scales were generally reliable and internally consistent. We developed scales of items from the surveys to help us understand whether our items were measuring the overall constructs that they were intended to measure. Altogether, we identified 27 scales in the student survey, 30 scales in the instructional staff survey, 28 scales in the school leader survey, and 20 scales in the district leader survey. We conducted CFA on scales from the student, instructional staff, and school leader surveys to assess whether items in scales were related. Because of limited sample size, we did not conduct CFA for the district leader survey. Because of the limited number of schools in our sample, all of our analyses were conducted on the individual level, and we do not make any claims about school-level phenomena based on our factor analyses.

> We did not find significant relationships between SCL and student achievement (as measured by the PSAT) in our Pilot Test II schools.

In general, respondents provided similar responses to scale items, and there was adequate evidence that items within scales were closely related. In a few cases, diagnostic information from the CFA analyses allowed us to remove problematic items or to gather more evidence when pairs of items correlated more strongly than could be explained by a single factor. In these cases, we made model revisions to improve the integrity of the scales. However, these changes were small and do not compromise the claims that there is evidence that items in a scale measure a common construct. In only three cases was there evidence that, even after modifications, scales did not perform adequately. The items from

Our analysis suggests that the MISCL Toolkit instruments may differentiate among levels of SCL in different schools.

those scales—all in the school leader survey—have been retained for individual item analyses.

We had evidence that, in general, the scales were internally consistent and could be used to distinguish among the responses of individual students, instructional staff, school leaders, and district leaders. In the few cases where reliability was low, the scales contained three or fewer items. Although including more items would make scales more reliable, it would also potentially increase respondent burden and jeopardize the quality of responses received. Importantly, the reported reliabilities here do not speak to whether schools can be distinguished based on the average responses of individuals associated with them. However, the inter-rater reliabilities estimated for the student survey offer some evidence that students share perceptions about school-level phenomena.

Relationships of SCL with external variables were not always consistent with theory. We used regressions to explore relationships between survey scales and outcomes that we hypothesized would be related, based on extant research. In the student survey, we examined whether responses to any of the 27 scales in the survey were associated with student outcomes and student and school characteristics. We hypothesized that more-extensive reports of SCL would be associated with higher student achievement and lower probability of suspension, and that higher-income and white students would report greater engagement in SCL than their lower-income and nonwhite counterparts. We distilled several key takeaways from this analysis:

- We did not find significant relationships between SCL and student achievement (as measured by the PSAT) in our Pilot Test II schools. We also did not find significant relationships between SCL and student background characteristics, including race/ethnicity and FRL eligibility. These results were not consistent with our expectations.
- Consistent with our expectations, students who reported higher SCL, according to some of our scales, were less likely to be suspended, although our sample size of suspended students was relatively small.
- Students in a school that described itself as experienced in SCL provided higher ratings of SCL than students in a school with less SCL experience, a result consistent with our expectations.

In the instructional staff survey, we examined whether responses to any of the 30 survey scales were associated with school characteristics. We specifically hypothesized that instructional staff who served more higher-income and white students (versus lower-income and nonwhite students) would report more-extensive use of SCL practices. Main findings from this analysis included the following:

- We found that instructional staff working in Title I schools (e.g., those receiving federal assistance because they serve high proportions of lower-income students) and instructional staff in schools with higher proportions of black and Hispanic students reported that they had inadequate resources to support SCL implementation.
- At the same time, we did not find any significant differences in instructional staff reports of SCL practices across the school-level characteristics that were available to us, including Title I schools and those serving more black and Hispanic students.
- Instructional staff at the Pilot Test II school that described itself as emphasizing SCL reported using SCL practices more extensively than instructional staff in the other school, where staff described themselves as newer to SCL.

In sum, the variety of evidence we collected suggested that the MISCL Toolkit measures aspects of

The evidence we collected suggests that the MISCL Toolkit measures SCL as intended, including some aspects of SCL that prior research suggests are related to positive student outcomes.

SCL as intended, including some aspects of SCL that prior research suggests are related to positive student outcomes. In addition, our analysis suggests that the MISCL Toolkit instruments may differentiate among levels of SCL in different schools. Specifically, students and instructional staff at the Pilot Test II school that described itself as emphasizing SCL reported using SCL practices more extensively than instructional staff and students in the other school, where staff described themselves as newer to SCL. Given that this finding is based on a sample size of two schools, much more research is necessary to determine whether the MISCL Toolkit can truly differentiate among schools with different levels of SCL. Lastly, we did not find clear and robust associations between SCL and other variables hypothesized to be related to SCL in our limited sample, which suggests that more research is necessary to understand these relationships. In particular, our results raise questions about whether SCL should always be expected to support improvements to student achievement or may be meeting other goals beyond those associated with achievement. However, as we discuss in the "Limitations" section later in this report, our small sample and limited data likely prevented us from uncovering key relationships between SCL and other variables.

To What Extent Was the MISCL Toolkit Usable and Useful to Those Who Undertook the Toolkit Process?

Toolkit users appeared to find the MISCL Toolkit process understandable and useful, although user concerns about the burden of Toolkit administration and perceptions of its evaluative nature led to revisions of some Toolkit content. Cognitive interviews and focus groups with Pilot Test II school leaders, instructional staff, and students who used the Toolkit suggested the following additional key takeaways regarding its usability:

- Although users found that the Toolkit process provided them with useful information, they also found the process burdensome and lengthy, which led us to reduce the complexity and length of our instruments and produce better summaries of our Toolkit process at the beginning of our User Guide.
- Some participants, including instructional staff and students, believed that the MISCL Toolkit was intended to be evaluative, which was not consistent with the intent to gather formative, constructive feedback. Therefore, we revised our User Guide and instruments to emphasize the need for clear communication about the purpose of the Toolkit and highlight that it should not be used for evaluation.
- Participants in the Toolkit process did not always appear to understand or fully engage in all aspects of MISCL Toolkit use. In particular, they did little to compare responses among different stakeholder groups (e.g., instructional staff and students) or compare subgroups within instructional staff or students, despite the Support Tools that were provided (i.e., planning worksheets, discussion guides, interactive data visualization and manipulation tool) in the MISCL Toolkit. This feedback led us to revise the User Guide to integrate a better

When creating any tool designed to support school practitioners in collecting data to reflect on and improve aspects of teaching and learning, developers face the tension between developing precise measures—which risk being extensive and complex—and measures that are productive and useful for practitioner reflection and improvement.

overview of the Toolkit process and guidance on reflection conversations and the importance of comparisons. The feedback also led us to develop an easily accessible, inexpensive tool for schools to collect, analyze and visualize survey data.

Limitations

Our study has several limitations. In particular, we fielded the entire MISCL Toolkit and collected validity evidence related to its implementation in two Pilot Test II high schools, although we also piloted the Toolkit surveys in two additional school contexts. Therefore, we cannot make generalizations about whether the Toolkit and instruments will be perceived and used in the same way in other school contexts. Nonetheless, the schools where we did collect data were quite different from one another and provided at least some validity evidence consistent with our hypotheses.

The limited sample sizes for these analyses also have implications for the kinds of validity and reliability evidence we were able to collect, which bear on the survey uses. Specifically, for the student survey and the instructional staff survey, we gathered evidence about individual perceptions, but sample size limitations (in terms of the number of schools represented in our samples) constrained our ability to gather evidence about the validity of inferences regarding school-level phenomena. For example, our data cannot tell us whether the typical SCL practices of schools were distinguishable based on the average responses of students or instructional staff. High internal consistency of instructional staff and student responses to items in scales intended to reflect particular SCL constructs does not automatically imply that school-level constructs are being measured precisely. The estimated regression relationships using the student survey suggest that its scales are limited in their capacity to describe school-level phenomena. These models employed school-fixed effects—therefore, we are unable to speak to the extent to which student perceptions, on average, are associated with school-level outcomes. Importantly, the fact that relationships exist within schools does not imply anything about the existence of relationships across schools. However, it is possible that these across-school relationships might differ in both magnitude and direction.

In addition, we caution that the MISCL Toolkit surveys, like any surveys, are subject to response biases. For example, students, instructional staff, or school and district leaders might have felt compelled to indicate that they were engaged in more SCL practices than they actually were using, given that their schools were focused on implementing SCL at some level. Alternatively, some respondents might have rated themselves as engaging in SCL at lower levels than they actually were using if they did not fully understand the survey item or were comparing themselves with peers or others who might be

providing SCL at higher levels. For these reasons, survey responses might not be the best way to assess the extent of SCL in high schools, although we hope that our cognitive interviews and other data we gathered in regard to response processes surfaced most of the issues related to response processes. Nonetheless, surveys were the main source of data for the SCL Toolkit because of the ease of fielding surveys over large populations and comparing results across students and respondents in different roles. To mitigate these inherent inaccuracies in survey data, we recommended that those using the MISCL Toolkit take advantage of the other instruments, including the classroom walkthrough and student focus group.

Implications for Development and Use of School-Based Tools for Measurement and School Improvement

Given the proliferation of tools and toolkits intended to help school systems and schools measure and improve their practices, we offer some lessons that may be relevant for practitioners using such tools and researchers developing similar tools. When creating any tool designed to support school practitioners in collecting data to reflect on and improve aspects of teaching and learning, developers face tension between developing precise measures—which risk being extensive and complex—and measures that are productive and useful for practitioner reflection and improvement. In addition, any set of measures or tools will have an optimal chance of being used productively if they are accompanied by simple and clear user materials that support the administration, analysis, and use of the data. Our experience developing the MISCL Toolkit, in conjunction with these considerations, led us to suggest five lessons for the development of tools intended to support school-based continuous improvement and the practitioners using those tools.

Lesson 1. Developers of data collection tools intended to support school improvement should start with a short and simple set of measures and instruments. We began by taking the results of our literature review—five SCL strategies and three contextual conditions—and breaking them down into multiple subcategories of specific instructional practices, which led to the creation of many items within our instruments. Eventually, we cut several items because they were redundant or closely related, and because—based on our pilot-testing—respondents found the instruments to be too long and burdensome. The result was a simpler set of measures and items. Therefore, toolkit developers might wish to start more simply and build in complexity as necessary, based on user feedback. This might also mean that those developing such toolkits might want to reduce the amount of ground and content they intend to cover through any single instrument given the burdens of using longer instruments.

Lesson 2. Developers of continuous improvement tools should provide a clear description of the capacity required to use those tools. As designed, the MISCL Toolkit requires a great deal of time and work to implement, particularly from the staff leading and facilitating the process. It involves arranging meetings to communicate the purpose of the Toolkit and describe the different activities, in addition to data collection from multiple sources. The MISCL Toolkit also requires time for analysis and, importantly, reflection on the data and what they suggest in terms of next steps for improvement. Without the will, commitment, and capacity to engage in these steps, schools might find the process challenging, unproductive, and even a waste of time for staff and students. Therefore, Toolkit developers should lay out the requirements from the outset to make clear what it will take for schools to use such tools and engage

Toolkit developers might wish to start more simply and build in complexity as necessary, based on user feedback.

Communication will ensure that those who are completing the instruments know why they are participating and feel confident that the data will be useful.

in an improvement process so schools can make an informed choice about whether to use them.

Lesson 3. An emphasis on open communication and use of a nonevaluative tone—by both developers of continuous improvement tools and users—can support buy-in and use of continuous improvement tools. We know from research that buy-in, trust, and engagement among all stakeholders—e.g., school leaders, instructional staff, students—is necessary for effective, sustained, and continuous improvement (Bryk and Schneider, 2002; Durlak and DuPre, 2008; Louis, 2007). Our MISCL Toolkit pilot tests suggest that buy-in, trust, and engagement were not necessarily secured when the district and school leaders facilitating Toolkit use were not explicit and frequent in their communication about the process. Therefore, we emphasized the need for such communication early in the User Guide and stressed clear communication about the goal and purpose of the MISCL Toolkit and instruments. We hope that communication will ensure that those who are completing the instruments know why they are participating and feel confident that the data will be useful. Those perceptions, in turn, can boost response rates to surveys and agreement to participate in other MISCL Toolkit activities, such as walkthroughs and reflection conversations.

Through our Toolkit pilots, we also learned that—even if facilitators do not intend for any data collected from teachers and students to be high stakes—participants may feel that they are being evaluated. Thus, we recommend that continuous improvement tools include regular reminders that such a process will work best if facilitators emphasize the nonevaluative, constructive, and information-gathering nature of those tools. Those messages could further help with buy-in and encourage respondents to provide authentic and honest answers when responding to surveys and participating in other activities. For example, some concepts in the MISCL Toolkit and survey instruments might not be well understood by school staff or students (particularly if the practices are not used in the school or if the school is in the early stages of exploring SCL). If staff do not feel they understand the concepts in the instruments, this could create self-consciousness, negative feelings, and concerns among respondents that their instruction is not student-centered "enough," which could lead to less authentic responses and decreased buy-in.

Lesson 4. Developers and users of continuous improvement tools should strive to create ample opportunities for participants to explore and compare data. The MISCL Toolkit included some opportunities for users to gather together and discuss data. However, the instruments provide a great deal of data, from survey responses of instructional staff, students, and leaders to qualitative data from walkthroughs and focus groups. Not only should these responses be considered, but—to get the most-useful data on equitable provision of SCL—users should be able to easily compare responses from different groups of instructional staff (e.g., instructional staff at different grade levels) and students (e.g., those of different grade levels, ethnicities, and genders). Such comparisons—between instructional staff in different grades and students of different income levels or ethnicities—could reveal potential inequities. However, in our pilot testing, users of our Toolkit made few efforts to compare responses among subgroups (e.g., students of different ethnicities and genders) or make comparisons among responses of instructional staff, students, and subgroups. Furthermore, when users did try to make these comparisons, they were hindered or confused by slightly different wording across surveys or different response scales. Therefore, we tried to improve mechanisms for users to examine data trends as easily as possible and compare different groups of

respondents, through revisions to survey items and our User Guide. We recommend that developers of toolkits and instruments like these devise ways for users to explore aggregated data individually and compare responses of different groups. In particular, tool developers should consider how to automate data collection through an online platform that supports and helps users make necessary comparisons. In addition, users of such tools should make an effort to help participants make those comparisons (e.g., by sharing the data well in advance of reflection conversations and encouraging participants to bring their own questions to the discussion).

Lesson 5. Developers and users of continuous improvement tools should involve students in continuous improvement efforts. The student survey and focus group protocols in the MISCL Toolkit provide important information about the extent to which students experience SCL practices and supports. Including responses from students—particularly in high school, where students could be expected to have well-developed opinions about their schools—is a key way to gauge the success of any set of practices or goals. In addition, the MISCL Toolkit suggests involving students in the reflections and discussions about data. According to our usability testing, students' perspectives are crucial to understanding the success of SCL implementation. Yet in testing our Toolkit, we observed less involvement of students than might be optimal. Of course, the extent of student involvement might depend somewhat on the age of students. That said, we recommend that developers of school-based continuous improvement tools think about how students can be involved in data collection, reflection conversations about the data, and plans for next steps. Furthermore, Toolkit facilitators and other adults should ideally solicit and encourage student feedback on a regular basis, because students might be hesitant to disagree and provide their valuable perspectives unless they are invited to do so.

Notes

[1] Appendix A is available online at www.rand.org/t/rr3235.

[2] The literature review is available online at studentsatthecenterhub.org/resource/measuring-scl-toolkit/.

[3] Appendix B is available online at www.rand.org/t/rr3235.

[4] For more information on these formulas, see Hu and Bentler (1999).

[5] Attendance data were not available from both schools and therefore not investigated in these analyses.

[6] The literature review is available online at studentsatthecenterhub.org/resource/measuring-scl-toolkit/.

[7] The other 69 studies were excluded based on varying criteria, including that they did not provide enough information about the SCL interventions under study or focused on aspects of teaching and learning not closely enough related to SCL.

[8] As noted in the introduction of the MISCL Toolkit, after we analyzed the data from this research, we decided to remove the teacher log from the Toolkit and recharacterize the Walkthrough Guide as an instrument. Therefore, the MISCL Toolkit now consists of six instruments, two Support Tools, and a User Guide.

References

Abakpa, Benjamin O., and Clement O. Iji, "Effect of Mastery Learning Approach on Senior Secondary School Students' Achievement in Geometry," *Journal of the Science Teachers' Association of Nigeria*, Vol. 46, No. 1, 2011.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 2014.

Austin, G. R., "Exemplary Schools and the Search for Effectiveness," *Educational Leadership*, Vol. 37, No. 1, 1979.

Awofala, Adeneye Olarewaju, "Effect of Personalisation of Instruction on Students' Motivation to Learn Mathematics Word Problems in Nigeria," *Turkish Journal of Computer and Mathematics Education*, Vol. 7, No. 3, 2016.

Billig, Shelley H., "Research on K-12 School-Based Service Learning: The Evidence Builds," *Phi Delta Kappan*, Vol. 81, No. 9, 2000, pp. 658–664.

———, "Does Quality Really Matter: Testing the New K–12 Service Learning Standards for Quality Practice," Barbara E. Moely, Shelley H. Billig, and Barbara A. Holland, eds., in *Advances in Service-Learning Research, Vol. 9: Creating Our Identities in Service-Learning and Community Engagement*, Greenwich, Conn.: Information Age Publishing, 2009, pp. 131–158.

Boulware-Gooden, Regina, Suzanne Carreker, Ann Thornhill, and R. Malatesha Joshi, "Instruction of Metacognitive Strategies Enhances Reading Comprehension and Vocabulary Achievement of Third-Grade Students," *The Reading Teacher*, Vol. 61, No. 1, 2007, pp. 70–77.

Brennan, Robert L., *Generalizability Theory*, New York: Springer, 2001.

Bryk, Anthony, and Barbara Schneider, *Trust in Schools: A Core Resource for Improvement*, New York: Russell Sage Foundation, 2002.

Cardelle-Elawar, Maria, "Effects of Teaching Metacognitive Skills to Students with Low Mathematics Ability," *Teaching and Teacher Education*, Vol. 8, No. 2, 1992.

Cavanaugh, Cathy, Kathy Jo Gillan, Jeff Kromrey, Melinda Hess, and Robert Blomeyer, *The Effects of Distance Education on K–12 Student Outcomes: A Meta-Analysis*, Naperville, Ill.: Learning Point Associates/North Central Regional Educational Laboratory (NCREL), October 2004.

Celio, Christine, Joseph Durlak, and Allison Dymnicki, "A Meta-Analysis of the Impact of Service-Learning on Students," *Journal of Experiential Education*, Vol. 34, No. 2, 2011.

Cole, Rachel, James J. Kemple, and Micha D. Segeritz, *Assessing the Early Impact of School of One: Evidence from Three School-Wide Pilots*, New York: New York University, Steinhardt School of Culture, Education, and Human Development, June 2012.

Conway, James M., Elise L. Amel, and Daniel P. Gerwien, "Teaching and Learning in the Social Context: A Meta-Analysis of Service Learning's Effects on Academic, Personal, Social, and Citizenship Outcomes," *Teaching of Psychology*, Vol. 36, 2009.

Cross, David R., and Scott G. Paris, "Developmental and Instructional Analyses of Children's Metacognition and Reading Comprehension," *Journal of Educational Psychology*, Vol. 80, No. 2, 1988.

Doyle, W., "Classroom Management Techniques," in O. Moles, ed., *Student Discipline Strategies: Research and Practice*, New York: State University of New York at Albany Press, 1990.

Dumas, J. S., and J. C. Redish, *A Practical Guide to Usability Testing*, Norwood, N.J.: Ablex Publishing Corporation, 1993.

Durlak, Joseph A., and Emily P. DuPre, "Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation," *American Journal of Community Psychology*, Vol. 41, No. 3–4, 2008, pp. 327–350.

Friedlaender, Diane, Dion Burns, Heather Lewis-Charp, Channa Mae Cook-Harvey, and Linda Darling-Hammond, *Student-Centered Schools: Closing the Opportunity Gap*, Stanford, Calif.: Stanford Center for Opportunity Policy in Education, June 2014.

Furco, Andrew, "Is Service-Learning Really Better Than Community Service? A Study of High School Service Program Outcomes," *Service Learning, General*, paper 154, 1996. As of January 30, 2020: https://digitalcommons.unomaha.edu/slceslgen/154

Gould, J. D., and C. Lewis, "Designing for Usability: Key Principles and What Designers Think," in R. Baecker, J. Grudin, W. Buxton, and S. Greenberg, eds., *Readings in Human-Computer Interaction, Toward the Year 2000*, New York: Morgan Kaufman, 1985, pp. 528–547.

Gross, Betheny, and Michael DeArmond, *Personalized Learning at a Crossroads: Early Lessons from the Next Generation Systems Initiative and the Regional Funds for Breakthrough Schools Initiative*, Seattle: Center on Reinventing Public Education, 2018.

Hu, L. T., and P. M. Bentler, "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 6, No. 1, 1999.

Kane, M., "The Errors of Our Ways," *Journal of Educational Measurement*, Vol. 48, No. 1, 2011.

Kaufman, Julia Heath, Rachana Seelam, Michelle W. Woodbridge, Lisa Sontag-Padilla, Karen Chan Osilla, and Bradley D. Stein, *Student Mental Health in California's K–12 Schools: School Principal Reports of Common Problems and Activities to Address Them*, Santa Monica, Calif.: RAND Corporation, RR-1129-CMHSA, 2015. As of January 30, 2020: https://www.rand.org/pubs/research_reports/RR1129.html

Kenny, D. A., B. Kaniskan, and D. B. McCoach, "The Performance of RMSEA in Models with Small Degrees of Freedom," *Sociological Methods & Research*, Vol. 44, No. 3, 2015.

Ku, Heng-Yu, and Howard J. Sullivan, "Student Performance and Attitudes Using Personalized Mathematics Instruction," *Educational Technology Research and Development*, Vol. 50, No. 1, March 2002.

Kulik, Chen-Lin C., James A. Kulik, and Robert L. Bangert-Drowns, "Effectiveness of Mastery Learning Programs: A Meta-Analysis," *Review of Educational Research*, Vol. 60, No. 2, 1990.

Kulik, James A., Chen-Lin C. Kulik, and Peter A. Cohen, "A Meta-Analysis of Outcome Studies of Keller's Personalized System of Instruction," *American Psychologist*, Vol. 34, No. 4, 1979.

LaBanca, Frank, Youn Joo Oh, Mhora Lorentson, Yueming Jia, Bernadette Sibuma, and Margot Snellback, *Blended Instruction: Measuring the Impact of Technology-Enhanced SCL on Academic Engagement, Skills Acquisition and Achievement of Underserved Students*, Quincy, Mass.: Nellie Mae Education Foundation, 2015.

Lewis, Matthew W., Rick Eden, Chandra Garber, Mollie Rudnick, Lucrecia Santibanez, and Tiffany Tsai, *Equity in Competency Education: Realizing the Potential, Overcoming the Obstacles, Competency Education Research Series*, Boston: Jobs for the Future, 2014.

López, Cecilia L., and Howard J. Sullivan, "Effect of Personalization of Instruction Context on the Achievement and Attitudes of Hispanic Students," *Education Technology Research and Development*, Vol. 40, No. 4, December 1992.

Louis, Karen Seashore, "Trust and Improvement in Schools," *Journal of Educational Change*, Vol. 8, No. 1, 2007.

Nunnally, Jum C., *Psychometric Theory*, 2nd ed., New York: McGraw-Hill, 1978.

Oakes, Jeannie, Tor Ormseth, Robert Bell, and Patricia Camp, *Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science*, Santa Monica, Calif.: RAND Corporation, R-3928-NSF, 1990. As of January 14, 2020: https://www.rand.org/pubs/reports/R3928.html

Pane, John F., Elizabeth D. Steiner, Matthew D. Baird, and Laura S. Hamilton, *Continued Progress: Promising Evidence on Personalized Learning*, Santa Monica, Calif.: RAND Corporation, RR-1365-BMGF, 2015. As of January 30, 2020: https://www.rand.org/pubs/research_reports/RR1365.html

Pane, John F., Elizabeth D. Steiner, Matthew D. Baird, Laura S. Hamilton, and Joseph D. Pane, *Informing Progress: Insights on Personalized Learning Implementation and Effects*, Santa Monica, Calif.: RAND Corporation, RR-2042-BMGF, 2017. As of January 30, 2020: https://www.rand.org/pubs/research_reports/RR2042.html

Patrick, Susan, Maria Worthen, Dale Frost, and Susan Gentz, *Promising State Policies for Personalized Learning*, Vienna, Va.: International Association for K-12 Online Learning, 2016.

Steele, Jennifer L., Matthew W. Lewis, Lucrecia Santibañez, Susannah Faxon-Mills, Mollie Rudnick, Brian M. Stecher, and Laura S. Hamilton, *Competency-Based Education in Three Pilot Programs: Examining Implementation and Outcomes*, Santa Monica, Calif.: RAND Corporation, RR-732-BMGF, 2014. As of January 30, 2020: https://www.rand.org/pubs/research_reports/RR732.html

Steiger, J. H., "Structural Model Evaluation and Modification: An Interval Estimation Approach," *Multivariate Behavioral Research*, Vol. 25, No. 2, 1990.

Steiner, Elizabeth D., Laura S. Hamilton, Laura Stelitano, and Mollie Rudnick, *Designing Innovative High Schools: Implementation of the Opportunity by Design Initiative After Two Years*, Santa Monica, Calif.: RAND Corporation, RR-2005-CCNY, 2017. As of January 30, 2020: https://www.rand.org/pubs/research_reports/RR2005.html

Steiner, Elizabeth D., Julia H. Kaufman, Elaine Wang, Karen Christianson, Laura S. Hamilton, and A. Ramos, "Measuring Student-Centered Learning Toolkit: Literature and Tool Review," Nellie Mae Education Foundation, undated.

Vandenberg, R. J., and C. E. Lance, "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," *Organizational Research Methods*, Vol. 3, No. 1, 2000.

Walkington, Candace A., "Using Adaptive Learning Technologies to Personalize Instruction to Student Interests: The Impact of Relevant Contexts on Performance and Learning Outcomes," *Journal of Educational Psychology*, Vol. 105, No. 4, 2013.

About the Authors

Julia H. Kaufman is a senior policy researcher at the RAND Corporation. Her research focuses on how states and school systems can support high-quality instruction and student learning, as well as methods for measuring educator perceptions and instruction.

Elizabeth D. Steiner is an associate policy researcher at the RAND Corporation with expertise in education policy, policy analysis, program evaluation, and qualitative methods and analysis. She is also a member of the Pardee RAND Graduate School faculty.

Jonathan Schweig is a social scientist at the RAND Corporation and professor at Pardee RAND Graduate School. His research focuses on education policy and the measurement of instructional practices, classroom and school climate, and social and emotional competencies

Sophie Meyers is a research assistant at the RAND Corporation. She works on education projects related to graduation pathways, student-centered learning, and other areas.

Karen Christianson is a policy analyst at the RAND Corporation. Her primary research interests concern efforts to eliminate racial and socioeconomic achievement gaps.

About This Report

This project and report are funded by the Nellie Mae Education Foundation (NMEF), the largest philanthropy in New England devoted completely to education. A key focus of NMEF's work with high schools is to implement a vision of student-centered learning (SCL) that focuses on the following four tenets: (1) learning is personalized; (2) learning is competency-based; (3) learning occurs anytime, anywhere; and (4) learning is student-owned. One way for NMEF to monitor and better understand the impact of its work is to develop data collection instruments that can measure the extent to which schools are implementing their vision for SCL.

NMEF asked the RAND Corporation to develop the Measuring and Improving Student-Centered Learning (MISCL) Toolkit for high schools that are interested in collecting, analyzing, and reflecting upon supports for SCL in their schools. NMEF also provided funding for RAND researchers to gather evidence of the validity, reliability, and usability of the instruments and User Guide that are part of the MISCL Toolkit. This report describes the process and the results of that work.

RAND Education and Labor

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. This study was sponsored by the Nellie Mae Education Foundation. For more information on the foundation, visit www.nmefoundation.org. More information about RAND can be found at www.rand.org. Questions about this report should be directed to Julia Kaufman at jkaufman@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

For more information on this publication, visit www.rand.org/t/RR3235.

Cover image: Fizkes/AdobeStock

www.rand.org

RR-3235-NMF