THE NATIONAL CENTER ON SCALING UP EFFECTIVE SCHOOLS

On the Development of Content-Specific Practical Measures Assessing Aspects of Instruction Associated with Student Learning

Nicholas M. Kochmanski | Erin C. Henrick | Paul A. Cobb |

Conference Paper
October 2015

VANDERBILT PEABODY COLLEGE

EDC Learning transforms lives. | Georgia State University | The University of North Carolina at Chapel Hill | Wisconsin

The National Center on Scaling Up Effective Schools (NCSU) is a national research and development center that focuses on identifying the combination of essential components and the programs, practices, processes and policies that make some high schools in large urban districts particularly effective with low income students, minority students, and English language learners. The Center's goal is to develop, implement, and test new processes that other districts will be able to use to scale up effective practices within the context of their own goals and unique circumstances. Led by Vanderbilt University's Peabody College, our partners include The University of North Carolina at Chapel Hill, Florida State University, the University of Wisconsin-Madison, Georgia State University, the University of California at Riverside, and the Education Development Center.

This paper was presented at NCSU's second national conference, Using Continuous Improvement to Integrating Design, Implementation, and Scale Up. The conference was held on October 7-9, 2015 in Nashville, TN. The authors are:

Nicholas M. Kochmanski
Erin C. Henrick
Paul A. Cobb
Vanderbilt University

This research was conducted with funding from the Institute of Education Sciences (R305C10023). The opinions expressed in this article are those of the authors and do not necessarily represent the views of the sponsor or the National Center on Scaling Up Effective Schools.

The analysis reported in this manuscript was supported by the National Science Foundation under grant DRL- 1119122. The opinions expressed in this paper do not necessarily reflect the views of the Foundation. The data that we present in this article are based on research conducted in collaboration with our colleagues from the University of Washington, Kara Jackson and Hannah Nieman.

Paper presented at the National Center on Scaling Up Effective Schools Conference in Nashville, TN on October 8th, 2015. This is a draft. Please do not cite without permission from the authors.

Abstract

This paper describes the development of *content-specific practical measures*, which we define as quick and accurate measures assessing aspects of the content-specific classroom learning environment associated with student learning. We present a set of five criteria to define content-specific practical measures in greater detail, and delineate a process for creating measures that meet these criteria. We illustrate each step in this process with examples from our current work developing content-specific practical measures assessing the quality of mathematics discourse.

Introduction

Educators and researchers continue to seek solutions to the challenge of improving instruction at scale. Instructional reforms that seek to "effect and sustain consequential change" often require multiple years and additional resources to achieve improvements (Coburn, 2003, p. 8). Efforts such as research practice partnerships (RPPs) and networked improvement communities (NICs) have begun to address some of the challenges posed by instructional improvement at scale by linking the improvement efforts of researchers and practitioners working on shared problems of practice (Penuel, Allen, Coburn, & Farrell, n.d.; Coburn, Penuel, & Geil, 2013; Bryk, Gomez, Grunow, & LeMahieu, 2015). When working together, researchers and practitioners can test and refine potential solutions to address an identified problem. Though these collaborative efforts appear promising, they are limited by a lack of available measures to quickly and accurately assess improvements to instruction. This poses a challenge for the effectiveness of these joint efforts, as practitioners and researchers. We agree with the tenant of improvement science that you can't improve at scale what you can't measure (Bryk et al, 2015).

Yeager, Bryk, Muhich, Hausman, and Morales (under review) address the need for quick and accurate measures to assess improvement by introducing the concept of *practical measures*, described as specific, targeted measures assessing elements closely associated with improvement work on relevant problems of practice. Inspired by Yeager et al., this paper describes the development of *content-specific practical measures*, which we define as **quick and accurate measures assessing aspects of the content-specific classroom learning environment**

associated with student learning. Developing content-specific practical measures that are quick and accurate could prove beneficial to efforts to support and understand improving instruction at scale.

Inspired by the work of the Carnegie Foundation for the Advancement of Teaching, a group of researchers and district leaders from multiple RPPs have been working to develop a common set of content-specific practical measures that could be used across districts and improvement efforts. This paper details an emerging process for the collaborative development of such measures. We believe the collaborative nature of this work (between researchers and district leaders across multiple contexts and partnerships) is essential to the process and products described within this paper, because it reflects the group's commitment to developing practical measures that are **not** boutique measures associated with one distinct improvement effort but are instead applicable across different school districts and research projects. Collaborating across diverse contexts and partnerships requires the development of shared goals that have relevance to all those involved.

This paper is organized around the following structure. First, we articulate a set of criteria for practical measures of instructional improvement. Next, we describe the process of developing these types of measures using illustrative examples from our current work and in doing so articulate the purpose and rationale for each step. Finally, we conclude by discussing both the next steps in this line of work and the implications for the use of practical measures to assess and support instructional improvement at scale.

Criteria for Content-Specific Practical Measures

It is important to clarify the distinction between practical measures, research measures, and accountability measures. Practical measures are quick, accurate measures that inform improvement efforts. Highly refined research measures are often developed to inform theory. Data from these measures are often not timely or specific enough to provide consistently beneficial, formative feedback to inform districts' instructional improvement efforts (Bryk et al, 2015).

Accountability measures also have limited usefulness as tools for instructional and programmatic feedback. Measures such as student achievement data from state-wide assessments are used to assess school or district progress towards accountability goals, but do nothing to highlight improvements to specific instructional practices. Other accountability measures, such as rubric-based teacher evaluation systems, require high levels of observational and instructional expertise (Riordan, Lacireno-Paquet, Shakman, Bocala, & Chang, 2015; Steinberg & Sartain, 2015), which limits the use of these rubric-based measures in instructional contexts without access to such expertise.

Drawing on the work of Yeager et al. (under review), we have developed the following criteria for content-specific practical measures of instructional improvement in the course of our current work.

1) The measures are explicitly linked to high-leverage, attainable improvement goals that are compelling to both practitioners and researchers. This first criterion emphasizes the critical link between improvement goals and measures. It is critical that the measure assess an aspect of the classroom learning environment that research has shown to be associated with student learning. Additionally, using practical measures to inform instruction is unlikely to be a meaningful

activity for practitioners unless the information provided is viewed as immediately relevant to their work.

2) The measures feature data collection and analysis routines that are relatively undemanding and can be used on a monthly, weekly, or even daily basis to provide prompt feedback and monitor progress. The second criterion focuses on the usability of the measures. We assume that the more the data collection processes leverage familiar tools and routines, the more likely that they will be used regularly. This criterion also points to the importance of developing, testing, and refining data collection and analysis routines in collaboration with district leaders, school leaders, and teachers to ensure these routines are feasible.

3) The measures orient educators to aspects of classroom learning environments associated with student learning, thereby serving as levers for as well as measures of improvement.

The third criterion highlights the dual purpose of the measures. In addition to assessing whether or not there has been improvement, the measures also have the potential to direct practitioners to aspects of the learning environment that may have been previously invisible to them. This can in turn orient practitioners to begin thinking about and developing practices that support student learning.

4) The measures highlight aspects of classroom learning environments that are potentially scalable across contexts and systems. The fourth criterion emphasizes that the measures developed are not boutique measures. They should focus on aspects of the classroom-learning environment that have a history of success across different contexts, and could thus prove useful to the field.

5) *The measures accurately assess observed elements of instruction, thereby producing data reflective of what happened in a classroom and/or learning context.* The fifth criteria highlights

the fact that content-specific practical measures should be sensitive to what occurs in the classroom, and should therefore result in data that consistently and accurately reports on what happened in a classroom or learning context.

These criteria are potentially useful for practitioners and researchers, as they help develop a shared understanding around content specific practical measures to improve the quality of instruction at scale.

A Process for the Development of Content-Specific Practical Measures

The work presented in this paper comes from a collaborative effort between researchers and district leaders to design a set of practical measures assessing aspects of the learning environment associated with student learning in upper elementary and middle-grades mathematics classrooms. The collaboration has focused on mathematics classrooms to-date because of a shared focus on goals for student learning in mathematics that are consistent with the newly adopted, more rigorous state standards (National Governors Association, 2010). These goals for student learning include the development of procedural fluency and deep conceptual understanding across mathematical domains. Attaining these learning goals requires a significant change in instructional practices, and this change requires significant learning for teachers (Franke, Kazemi & Battey, 2007).

Throughout this design work, we have stepped back from the process of developing measures specific to mathematics classrooms to delineate a more general process for creating such measures that meet the above criteria. In this section of the paper, we describe this process and use illustrative examples from our current work. Although the illustrations focus on mathematics instruction, the steps we outline are relevant to the development of content-specific practical measures regardless of the content area.

Step 1: Identify a shared improvement focus

The first step of the development process is to *identify a shared improvement focus*. This step involves bridging the diverse needs and goals of researchers and practitioners to determine an improvement focus that is mutually beneficial and feasible. Determining an improvement focus entails a series of discussions about potential improvements in instructional practices that would be of interest to the participating researchers and also of benefit to practitioners. We suggest beginning this step with a brainstorming session with researchers and practitioners to select an improvement focus that 1) fits within the current district improvement initiatives and 2) is currently hampered by a lack of instrumentation to quickly and accurately assess current and future practices.

In the context of our collaborative effort, the *quality of classroom discussions* emerged as the focal area for our joint work, in part because a significant body of research has generated a robust understanding of what constitutes high-quality discourse in mathematics classrooms, and why it is crucial if students are to develop both conceptual understanding and procedural fluency (Franke, Kazemi, & Battey, 2007). For example, it is critical that students are both pressed and supported to explain and justify their reasoning in ways that other students can understand (Kazemi & Stipek, 2001; Wood, Cobb, & Yackel, 1993). This emphasis on the quality of student discourse is apparent in the Common Core State Standards- Mathematics (CCSS-M) Practice Standards, which emphasizes the need for students to engage in disciplinary practices such as justifying solutions, constructing viable arguments, and critiquing the reasoning of others. Additionally, we knew that the majority of the classrooms in our partner districts featured either no discussion of students' mathematical reasoning or discussions in which students shared their answers in sequence, with little to no genuine discourse. For our partner districts, improving the

frequency and quality of classroom discussions would thus mark a significant improvement to the quality of mathematics instruction and to students' opportunities for learning in mathematics classrooms.

We collaboratively developed an improvement goal that "students will explain their mathematical thinking in ways that other students can understand." With this shared improvement goal in mind, our next step was to consider the system of district and school level supports that have the potential to improve the quality of student explanations and classroom discussions.

Step 2: Map the System of Improvement Supports Linked to the Improvement Goal

Influenced by improvement science, *the second step in the process is to map the system of supports that have the potential to positively impact the improvement goal.* This activity involves identifying the many different parts of the system that can potentially impact the improvement goal. We agree with Bryk et al. that "adopting a systems perspectives makes visible many of the hidden complexities actually operating in an organization that might be important targets for change" (p. 14).

Engaging in this type of mapping activity is useful for researchers and practitioners seeking to develop content-specific practical measures. First, it helps to develop a context for the core problem of practice and encourages making sense of and accounting for the relationship between the many factors contributing to the core problem of practice. Second, it creates a common language and understanding between participating groups and stakeholders. Third, it provides an opportunity to begin developing a "practical" theory for what has contributed to the current state of the system, and what elements would support improvement to the system.

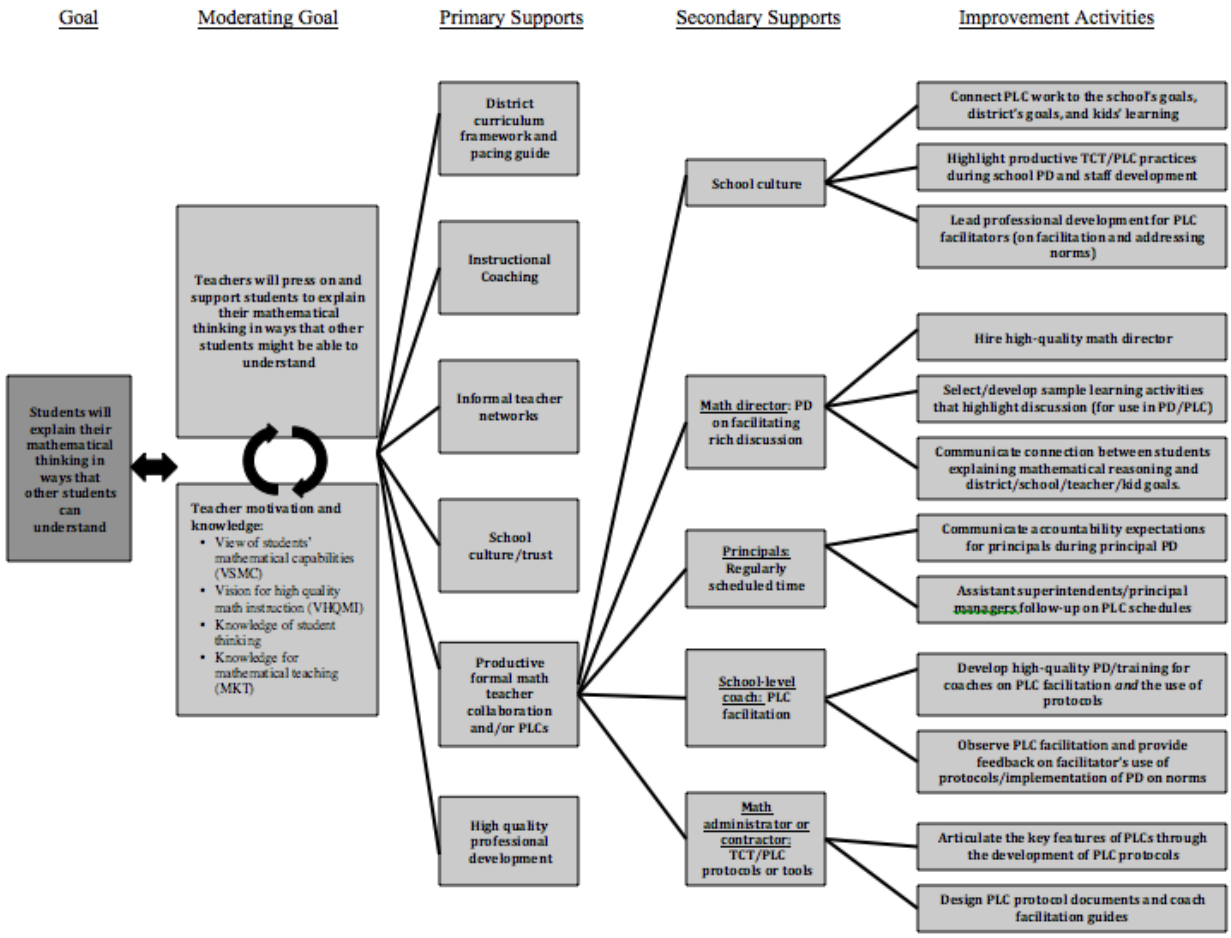
Finally, mapping the system helps identify additional focus points for the further development of practical measures specific to the problem of practice.

In the context of our work, we developed an understanding of the system of organizational supports that might influence student discourse. With our RPP partners, we developed an initial "diagram of supports" modeled after the improvement science driver diagram to represent the major supports that would lead to improvement towards this goal (see figure). We conjectured that in order for students to explain their mathematical thinking in ways that other students can understand, teachers needed to press and support students to develop this practice in their classrooms. When designing a system of supports to improve this goal, we identified five different primary supports that districts and schools could provide and improve upon to improve the nature of student explanations. These included a district curriculum framework and pacing guide, informal teachers networks, school culture and trust, productive formal math teacher collaboration, and high quality professional development. To further elaborate our understanding, we then developed a group of secondary supports related to each primary support. For example, both districts determined that providing time for teachers to collaborate was a key support to improving teachers' instruction. We conjectured that there were four key areas of support that could improve the quality of teacher collaborative time (TCT) in a way that would support teachers' ability to press students to explain their mathematical thinking. These included 1) district PD on facilitating discussion and how to organize TCT time at the school level to improve the quality of discussion, 2) regularly scheduled time for teachers to meet, 3) an expert TCT facilitator, and 4) TCT protocols or tools.

Finally, for each of the secondary supports, we articulated potential for improvement activities that we could enact within the system. We could then test to see if the enactment of

these improvement activities would result result in an improvement on our goal. For example, if we want to improve the effectiveness of the TCT facilitator, we could support the TCT facilitator through professional development on facilitating effective teacher meetings on improving the quality of discourse in mathematics classrooms.

Figure 1: Map of System Supports



Step 3: Conduct Literature Scan and Select an Appropriate Type of Measurement Instrument

The third step in the process to develop practical measures for instructional improvement is to *conduct a literature scan on the identified improvement focus and select an appropriate measurement instrument*. The purpose of the literature review is to 1) understand current findings related to the identified improvement focus and 2) review currently available measures related to the identified improvement focus. Conducting a scan of available measures related to the improvement area orients the designer toward measurement instruments with the potential to reliably assess the focal topic.

Our literature scan indicated the importance of distinguishing between classroom discussion that is calculational or conceptual in nature. Cobb, Gresalfi, and Hodge (2008) describe explanations that are calculational in nature as "specifying the calculational steps taken to produce a result" (p. 15). In classrooms featuring discussions that are calculational in nature, students are more likely to "describe the steps they took to solve a problem without explaining why the solution works mathematically" (Kazemi & Stipek, 2001, p. 64). This is different from classrooms featuring discussions that are conceptual in nature, as these classrooms feature not only explanations of how a student produced a result, but also a justification for "why a solution gave insight into the question at hand" (Cobb, Gresalfi, and Hodge, 2008, p. 16). This indicates that what counts as an acceptable mathematical explanation is often different depending on the nature of the discussion, which is a critical distinction, because students are more likely to understand one another when they engage in conceptual discussions (Cobb, Gresalfi, and Hodge, 2008; Cobb, Stephan, McClain, & Gravemeijer, 2001)

It is also critical to *select a type of measurement instrument that fits the practical measures criteria.* There are compromises that occur when trying to determine the best type of measure that meets the criteria and provides valuable information related to the core

improvement focus. For example, while classroom observation and using a rubric based assessment would perhaps provide a deeper understanding of aspects of classroom discussion, this would not meet the practical measurement criteria of a simple and feasible data collection and analysis process that could be scaled across a district. Below are questions, related to our practical measures criteria, which proved useful when thinking about the type of measure:

- Can a teacher or school based personnel collect and analyze this data quickly and easily given time and resource constraints?
- Would it require limited training to use the measure?
- Is this scalable, can this be used across a system?
- Does this measure have the potential to have face-value with teachers?

Based on these questions, and in discussion with district level mathematics leaders, we determined that quick, 1-3 minute student surveys would be the most appropriate measurement instrument for measuring the quality of student discourse in mathematics classrooms that could be implemented and analyzed across a large urban district. Student surveys offer greater insight into the classroom environment than teacher perception surveys, as students are more likely to accurately report what occurs in the classroom (Bill & Melinda Gates Foundation, 2013). As a measurement instrument, student surveys also require minimal training to administer, and can be analyzed quickly. To inform the structure of our student survey questions, we also examined several different student perception surveys, including those described in the Measures of Effective Teaching study (Bill & Melinda Gates Foundation, 2013).

We then scanned the literature to see if there were any student surveys designed to assess the quality of mathematics classroom discourse and did not find any. We then examined two research measures assessing aspects of discourse in mathematics classrooms to help us develop

our own student survey questions: the Instructional Quality Assessment (Boston and Wolf, 2006) and the TruMath rubrics (Schoenfeld, 2014). Although these measures did not meet the criteria for a content-specific practical measure (in part because they require significant training to establish reliability across coders and cannot be collected and analyzed in a short time period), examining them informed our understanding of key aspects of student discourse specific to mathematics classrooms.

Our scan of the literature on discourse in mathematics classrooms—in combination with our review of mathematics rubrics and student perception survey items—informed the development of our initial student survey items on classroom discourse. We targeted aspects of the instructional environment, including: 1) whether students understood other students' explanations of their mathematical reasoning and 2) aspects of classroom discussion linked to student understanding and learning. For example, we knew from our review of the literature on student discussion that students are more likely to understand discussion that is conceptual in nature, so we sought to write student survey questions assessing this construct. When writing the survey items, we used language that we thought would communicate to upper elementary and middle-grades students.

Step 4: Design and Improve a Specific Measure through Cycles of Trial, Analysis, and Revision

The fourth step in the development process is to *design and improve a specific measure through cycles of trial, analysis, and revision*. This fourth step follows from the tradition of design research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) and features iterations of a modified cycle of design and analysis. The modified cycle begins with a trial of potential measures in authentic, diverse instructional contexts and is then followed by a formal analysis of

the data collected from the trial. This formal analysis of the data informs the final phase in the cycle, which focuses on empirically-supported revisions to the initial measures. These revised items are then the basis for subsequent trials.

Phase 1 of the Cycle: Trial Measurement Items in Authentic, Diverse Instructional Contexts

The first phase in the cycle entails *trialing the measurement items in authentic, diverse instructional contexts*. Trialing potential measures involves collecting data on the use and interpretation of the measures that, when formally analyzed, could directly inform potential revisions. There are two key considerations when trialing potential measures. First, it is important to select a context that is authentic to the problem of practice addressed by the improvement focus and also diverse enough to allow for analyses comparing the performance of the measures across a range of instructional environments (e.g. across classrooms of different quality, across classrooms featuring different ranges of student populations, etc.). Different contexts afford different opportunities to ensures the measures communicate consistently regardless of the type of learning environment or the quality of instruction. This also ensures the measures are not boutique measures and have the potential to inform varied and diverse improvement efforts.

Our trial of the classroom discussion surveys provides an illustration of the importance of selecting a diverse, authentic context for the trial of potential measures. In conjunction with our district partners, we decided to trial our potential measures in middle school and upper-elementary classrooms identified by our district partners as featuring either calculational or conceptual discourse. As described earlier, this distinction was important for our collaboration, because the nature of the learning environment might impact students' responses to questions about whether or not they understood their peers' explanations, as students are more likely to

make sense of conceptually-grounded mathematical explanations (Cobb, Stephan, McClain, & Gravemeijer, 2001). Trialing our measures in classrooms featuring either calculational or conceptual discourse was thus essential to testing whether our potential measurement items communicated to students regardless of the learning environment.

Second, it is important to collect different forms of data to from each trial. Two forms of data that are especially beneficial are 1) observations of the instructional environment, as these observations foster comparisons between what is observed and what the measures assess, and 2) cognitive interviews (Desimone & LeFloch, 2004) with stakeholders, which support productive revision by producing data on how stakeholders interpret the measurement items. These different forms of data can then inform subsequent steps in the development process by providing cross-checks for the performance of different items, thereby indicating whether or not items are problematic or productive. For example, comparing data from classrooms (such as observations of student discourse or field notes from the classroom) with data from the measures (in our case, student response data) allows for the analysis of whether or not the measures are assessing what occurred in the classroom. Collecting different forms of data can also provide additional ideas for the revision of problematic items. For example, analyzing data from cognitive interviews with stakeholders can identify instances where the measures do not communicate consistently across users.

Collecting different forms of data was an especially salient feature of our trials for the development of student surveys of mathematics discourse, because we were unsure whether our surveys were sensitive to the nuances of classroom discourse. In order to test this, we linked trials of the student survey questionnaires with observations of classroom discussions and cognitive interviews with students. While the observations provided us with a comparison point

for student responses on the surveys, the cognitive interviews allowed us to probe on students' interpretations of individual questions. For example, we asked students to explain their interpretation of a potentially problematic question, "I understood how other students in my group found answers." We probed on students' interpretations of "understood" and "found answers" because we were unsure of whether this language would communicate consistently across students and classrooms. Collecting observational data alongside interview data with students allowed us to test whether the student surveys accurately assessed what happened in the classroom, and informed revisions to problematic questions.

Phase 2 of the Cycle: Analyze Resulting Data to Assess the Usefulness of the Measurement Items

After trialing the items, the next phase in the cycle is to *analyze the data to assess the usefulness of the measurement items* across responses and contexts. This phase involves conducting a formal analysis of whether or not the potential measures meet the criteria for content-specific practical measures, as well as any context-specific or goal-specific criteria for useful measures. Consistently meeting these these requirements signifies that the measures are useful. The data collected from this analysis can then inform subsequent revisions of the measurement items that show promise.

In the case of classroom discussions, we determined early on to focus the bulk of our analysis on the fifth criterion for content-specific practical measures, which states that measures should accurately assess observed elements of instruction, thereby producing data reflective of what happened in a classroom. This was important because we were unsure whether our questions would be sensitive to key distinctions in the quality of mathematics discourse. To this end, we sought to determine which measurement items were proving effective and which items

needed substantive revisions. This entailed comparing trends in student responses to our survey questions with our observations of classrooms, identifying trends in student response data based on whether the classroom discourse was conceptually-oriented or calculational in nature, and then analyzing audio recordings of our cognitive interviews with selected students to understand why individual measurement items did not communicate to students. This data thus informed our eventual revision efforts.

One of our initial student survey questions, "I understood how other students in my group found answers" provides a particularly strong illustration for how the analysis phase can inform subsequent revisions. By analyzing our available data, we discovered that student responses to this question did not align with what we observed in classrooms. Students participating in classrooms featuring conceptually-oriented discussions typically responded either agree or strongly agree to this question, which confirms what we observed in the classrooms—that students enagaging in conceptually-oriented discussions were explaining their thinking clearly and were engaging in strong conversations with their peers around different ways to solve problems. A high percentage of students in classrooms featuring calculational discussions also strongly agreed with the statement, despite the fact that we observed discussions that predominantly focused on sharing answers or pointing our small errors in calculations. While students in classrooms featuring calculational discussions did indicate a wider range of responses to this question, the number of students claiming to understand how other students in their group found answers did not appear to match our initial observations. Table 2 outlines the spread of student responses below.

Table 1: Spread of Responses to Student Survey Item #4

4) I understood how other students in my group found answers. (Choose one.)

Classroom Type	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
Calculational	0%	27%	45%	27%
Conceptual	0	0	77%	22%

Because the student responses to this question ran contrary to many of our observations of the classroom environment, we decided to compare this data against our cognitive interviews with students. This helped us determine whether the large number of "strongly agree" responses were a result of students' interpretations of the questions or something else. Analyzing the cognitive interview audio in conjunction with the data from students' responses indicated that students in classrooms featuring calculational discussion who understood how other students in their group found answers because they solved the problems the *same way* as their partners. For these students, the word "understand" had a different meaning than students in more conceptual classrooms and was essentially synonymous with "getting the right answer."

This distinction was particularly clear in one student's description of his rationale for choosing "agree" to this question:

> *"If they sort of had the same idea, then I could sort of comprehend of what [the other student] was getting to, if [the two students] had the same idea… If they both had the exact same thing, um, written I could put strongly agree because that could help me understand way a lot better, but since one didn't put, umm, as much as [the other] then I can't really put that."*

For this student, the fact that he and his partner did not have the exact "same idea" kept him from selecting "strongly agree." In classrooms featuring calculational discussion, students often use the same strategy to solve problems, which might in turn explain why many students in these classrooms selected strongly agreed, despite the fact that students' explanations often amounted to sharing answers—they had the same method, so when another student shared his or her answer, the student treated this as a confirming explanation.

Further analysis of students' interpretations of this question indicated that students also had inconsistent interpretations of the very meaning behind this question. For example, one student explained his answer to this question by stating:

> *"When [one student] described his [method], he had very little to say about it, so it was a little more confusing about what to do. But when [another student] did it, she had a lot down of what she did, and so it helped me understand a little bit better."*

When pressed on what he meant by "a lot down," the student explained:

> *"A lot down… A lot down… She had a lot of words on her paper, and so, for me, more helps me understand better—less doesn't make more sense to me. So if there is more, to me, it's better."*

For this student, the idea of understanding another student's explanation was predicated on the *length* of the explanation. Another student saw this question differently, noting:

> *"You get the concept of—either what the person is telling you or you get the concept of the problem. So you, you, you almost know how to digest it in your brain."*

Here the student explained "understood" as relating to the "concept." For this student, understanding an explanation depended on whether or not she could make sense of the other students' concept of the problem. These different interpretations of the core language of the

survey question highlighted another element of confusion within this question, which may have further impacted the way students responded.

Our analysis of the cognitive interviews thus highlighted two key problems in the question, "I understood how other students in my group found answers." First, students in more procedurally oriented classrooms were more likely to solve problems the same way, which may have accounted for the large number of students who responded positively to this question. This was in spite of the fact that the explanations in the classrooms featuring calculational discussions were often underdeveloped and/or relied on sharing answers explicitly. (For example, an explanation might consist simply of: "I got 32 for that one.") Second, the language of the question was confusing to students, as different children interpreted different words (e.g., X, Y) within the question in different ways.

This data informed our revisions to the potential measurement item. By comparing the student responses to our observations of the classroom event, we could determine areas where the student survey did not measure the same thing as what we saw in the classroom. This incongruity suggested the question needed refinement before its use in subsequent iterations of the cycle. We could then use the cognitive interview data to identify what was problematic for specific survey items and address these issues in our revision of the items.

Phase 3 of the Cycle: Revise the Measurement Items for Further Iterations of the Cycle

The third and final phase in the cycle involves *revising the measurement items to support further iterations of the cycle*. This typically means adjusting the measurement items in an effort to develop a more useful version of the measure. We engaged in this third phase of the cycle by altering the language of many of our questions in response to the trends we discovered while analyzing our data. We also removed some survey questions that were either repetitive or did not

communicate to students. The student survey question, "I understood how other students in my group found answers," is illustrative of this process, as it underwent a number of revisions following the analysis described above. These revisions resulted in the following wording: "Did talking to other students in your group help you understand their thinking?" We shifted the language of the question to reflect "talking" as the core action assessed, as opposed to whether or not a student "understood" another student. Though our subsequent question still included the word "understanding," the new wording shifted the focus from whether or not the student "understood" something to whether or not talking to other students was helpful in understanding their thinking. This shifts the focus away from different conceptions of what it meant to understand, which we conjecture vary between classrooms featuring conceptual and calculational discourse. Instead, focusing on talking as a *mechanism* for understanding would foster responses more consistent with our observations of the classroom event. Though we acknowledge that this question may require additional revision, the aforementioned changes may result in student responses that indicate whether or not students were able to understand others' thinking, as opposed to the way to find an answer.

Further Iterations of the Cycle

It is unlikely that the development of content-specific practical measures will require only one cycle of trial, analysis, and revision, because the very nature of the cycle fosters revisions to the measures. These revisions in turn require trials in authentic, diverse contexts, analysis of the trial data, and further revision. These iterations of the cycle should then continue until the measurement items consistently meet each of the criteria for content-specific practical measures.

Our cycles of trial, analysis, and revision for the student surveys have focused so far on the fifth criterion for content-specific practical measures: the measures accurately assess

observed elements of instruction, thereby producing data reflective of what happened in a classroom and/or learning context. As our work progresses, we expect to continue to engage in iterations of this cycle until we have developed measures that consistently meet all of the criteria. Once we have done so, we expect to move into the final step in the development process.

Step 5: Collaborate with Stakeholders to Develop Routines for Using of the Measure

The final step involves *collaborating with stakeholders to develop routines for using the measure.* We define stakeholders as practitioners (such as teachers, school leaders, and district leaders) and researchers who intend to use the measures to assess and leverage instructional improvement. Though we have yet to engage in this final step, we anticipate that it will involve collaborating with a diverse and representative group of stakeholders to develop routines for the administration and use of the practical measures. We anticipate that developing these routines will entail: 1) consulting relevant teacher groups for feedback on the administration, use, and relevance of the measures; 2) co-constructing potential data displays or dashboards; and 3) constructing tools and protocols to support the discussion and analysis of the data gleaned from the practical measures. This final step is important because it ensures stakeholders working in different instructional contexts—with different capacities for administration, data analysis, and interpretation—can use the measures. By working collaboratively with stakeholders, we can ensure that the user remains at the forefront of the measures, thereby addressing many of the criteria for content-specific practical measures, including criterion two: the measures feature data collection and analysis routines that are relatively undemanding and can be used on a monthly, weekly, or even daily basis to provide prompt feedback and monitor progress.

For our work with student surveys, we first plan to determine the face validity and perceived use of the surveys by interviewing teachers, school leaders, and district leaders. We

expect that these interviews will probe on teachers' interpretation of the questions and how the questions could provide an opportunity for them to reflect upon (and orient them toward) more ambitious instructional practice. For example, the question, "Did talking to other students in your group helped you understand their thinking?" may prompt teachers to discuss which practices could support students to better understand their peers during group work or whole-class discussions with a coach, administrator, or lead teacher. We could then see these conversations as orienting teachers to more sophisticated instructional practices that may support the development of more productive discourse in the classroom.

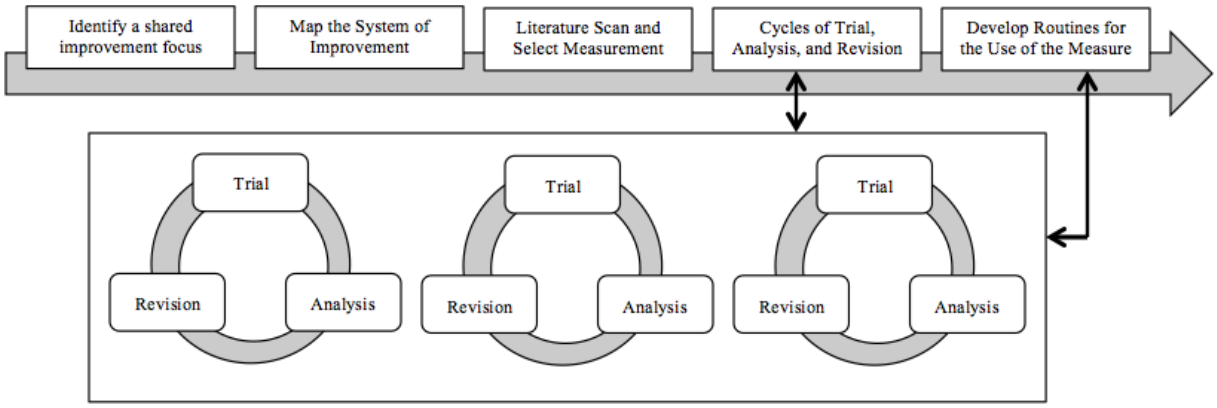
We also intend to collaborate with district leaders, school leaders, and teachers to ensure that the processes for the administration of the student surveys are simple and user-centric. Options for the administration of the surveys may include online surveys that students could take on iPads, paper and pencil survey for those teachers and districts without access to such forms of technology, or surveys specifically developed for in-class technologies, such as clickers or networked calculators. The development of routines or heuristics for the interpretation of the student surveys could facilitate the integration of these measures into everyday practice, thereby reinforcing the formative nature of the measures, and further distancing the student surveys from accountability measures.

Implications and Next Steps

Each of the steps above contributes to the development of content-specific practical measures that assess aspects of instruction associated with student learning. Taken together, the development steps articulate a *process for the development of content-specific practical measures* that we believe could be beneficial to others interested in developing such measures,

regardless of subject or instructional focus. We see each of the steps in the process as building on the previous steps, as illustrated by the process diagram below.

Diagram 2: Process Diagram for the Development of Measures



The first step, to identify a shared improvement focus, calls for the articulation of an improvement focus that is relevant for all participating practitioners and researchers. The second step, mapping the system of improvement supports that contributes to this improvement focus, then situates the focus within a broader program of work. This situated perspective then informs the selection of an appropriate measurement instrument by indicating what form of measure is most appropriate for the given context. Selecting a measurement instrument that is appropriate for the needs of the improvement focus can help to ensure the content-specific practical measure does not overburden potential users. Cross-checking potential measurement instruments against

the literature on instructional improvement, as well as measurement, ensures potential measures are appropriate to the improvement system and also supported by what is known in the literature.

The fourth step of the development process involves engaging in cycles of trial, analysis, and revision. These iterative cycles can involve an extensive period of work, as the purpose of each cycle is to improve and revise the measures. The process for developing content-specific practical measures is then followed by the final step, to gather feedback from stakeholders to develop routines for use of the measure. This final step ensures the measures are useful for multiple stakeholders within the instructional system.

When read left to right, the diagram presents a generalizable process for the development of content-specific practical measures. This generalizable process appears to be applicable across content areas and grade levels. Additionally, we see these steps as general enough to support the development of content-specific practical measures regardless of the type of measure, and is not limited to survey-based measures.

Developing a system of practical measures to support instructional improvement at scale

We anticipate extending our focus to other high-leverage aspects of classroom instruction that have been linked to student learning in mathematics and that attend to issues of equity. For example, we are considering developing a practical measure of the extent to which teachers introduce rigorous mathematical tasks in ways that support all students' substantial participation (Jackson et al., 2013).

Additionally, we are considering creating a *system of measures* (Bryk et al., 2015) that will connect measures of the quality of instruction (e.g., the quality of discourse) to measures of the quality of supports and press for instructional improvement. This collection of measures could potentially inform the improvement work of districts by providing ways to assess the

quality of supports offered to teachers as they work to improve identified instructional practices. For example, we intend to develop a practical measure to assess the quality of teacher collaborative meetings in which teachers share and discuss practical measures data from their classrooms related to the quality of classroom discourse. In these meetings, teachers could discuss the possible reasons for students' responses indicating low-quality of classroom discourse, which could include reasons such as using low level tasks, not pressing students to explain their reasoning, or telling students the procedure to solve the problem. Based on their diagnoses of the problem, teacher could then discuss strategies for improving the quality of discussion in future lessons. In this scenario, one could imagine the district providing professional development to teachers around improving student discourse and useful activities to support this work in teacher collaborative meetings.

It is important to note that using a system of practical measures in this manner would entail expanding the intended users to include school leadership teams, which would require attending to administrators' capacity to use such data to make informed decisions about how to target supports for teachers.

In conclusion, we believe that a system of measures could be very useful as it would allow researchers and practitioners the ability to work together to differentiate between limitations of implementation and limitations of the practical theory that underpins the improvement effort. This would be a very important step forward as researchers and practitioners collaborate to improve the quality of instruction and student learning opportunities across a school district. The creation of viable systems of measures would also enable researchers to conduct formal analyses of the relation between improvements in supports and

press for teachers' improvement of their instructional practices, changes in their actual practice, and changes in student learning.

WORKS CITED

Bill & Melinda Gates Foundation. (2013). *Nine Principles for Using Measures of Effective Teaching*. Retrieved from www.metproject.org.

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of the instructional quality assessment toolkit. CSE Technical Report 672* (No. 672). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Cobb, P., Gresalfi, M., & Hodge, L. L. (2008). An Interpretive Scheme for Analyzing the Identities That Students Develop in Mathematics Classrooms. *Journal for Research in Mathematics Education, 39*, 1–29.

Cobb, P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in classroom mathematical practices. *The Journal of the Learning Sciences, 10*(1 & 2), 113-163.

Coburn, C. E. (2003). Rethinking Scale: Moving beyond Numbers to a Deep and Lasting Change. *Educational Researcher, 32*(6), 3-12.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (January 2013). *Research Practice Partnerships: A Strategy for Leveraging Research for Educational Improvement in School Districts*. William T. Grant Foundation, New York, NY.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational evaluation and policy analysis, 26*(1), 1-22.

Elias, M. J., Zins, J. E., Graczyk, P. A., & Weissberg, R. P. (2003). Implementation, Sustainability, and Scaling Up of Social-Emotional and Academic Innovations in Public Schools. *School Psychology Review, 32*(3), 303–319.

Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. *Second handbook of research on mathematics teaching and learning, 1*, 225-256.

Hatch, T. (2000). What Does it Take to "Go to Scale"? Reflections on the Promise and the Perils of Comprehensive School Reform. *Journal of Education for Students Placed at Risk (JESPAR), 5*(4), 339–354.

Kazemi, E., & Stipek, D. (2001). Promoting Conceptual Thinking in Four Upper Elementary Mathematics Classrooms. *The Elementary School Journal, 102*(1), 59–80.

Marrongelle, K., Sztajn, P., & Smith, M. (2013). Scaling Up Professional Development in an Era of Common State Standards. *Journal of Teacher Education*, *64*(3), 202–211.

National Governors Association, Common Core State Standards Initiative (2011). *Common Core State Standards in Mathematics.* Retrieved July 5, 2015, from: http://www.corestandards.org/Math.

Penuel, W. R., Allen, A. R., Coburn, C. E., & Farrell, C. (2015). Conceptualizing research–practice partnerships as joint work at boundaries. Journal of Education for Students Placed at Risk (JESPAR), 20(1-2), 182-197.

Riordan, J., Lacireno-Paquet, N., Shakman, K., Bocala, C., & Chang, Q. (2015). *Redesigning teacher evaluation: Lessons from a pilot implementation* (REL). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Schoenfeld, A. H. (2014, November). What makes for powerful classrooms, and how can we support teachers in creating them?*Educational Researcher*, 43(8), 404-412.

Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In Lester, F. K. (Ed.), in *Second Handbook of Research on Mathematics Teaching and Learning* (319-369). Charlotte, NC: Information Age Publishing.

Steinberg, M. P., & Sartain, L. (2015). Does better observation make better teachers? *Education Next*, *15*(1).

Cobb, P., Wood, T., & Yackel, E. (1993). Discourse, mathematical thinking, and classroom practice. Contexts for learning: Sociocultural dynamics in children's development, 91-119.

Yeager, D. S., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (under review). Practical Measurement.