AERA Open
April-June 2016, Vol. 2, No. 2, pp. 1–13
DOI: 10.1177/2332858416643077

Why Don't Teachers Understand Our Questions? Reconceptualizing Teachers' "Misinterpretation" of Survey Items

Anne Garrison Wilhelm
Southern Methodist University

Christine Andrews-Larson
Florida State University

This study examined sources of inconsistency between teachers' and researchers' interpretations of survey items. We analyzed cognitive interview data from 12 middle school mathematics teachers to understand their interpretations of survey items focused on one aspect of their practice: the content of their advice-seeking interactions. Through this analysis we found that previously documented conceptualizations of sources of misinterpretation within teacher surveys (e.g., structural complexity, use of reform language) did not adequately account for all of the inconsistencies between the survey items and teachers' interpretations. We found it useful to reconceptualize the broader source of many of the misinterpretations as an issue of fit between the researchers' intended interpretation and teachers' professional practice.

Keywords: *cognitive interviews, teacher practice, communities of practice, survey research*

The advancement of science relies on the collection and analysis of data that accurately capture phenomena of interest. In the context of education, observations and interviews can function as rich sources of data for documenting and analyzing ways to understand and improve various aspects of teaching and learning. However, these forms of data tend to be time-consuming and expensive to collect and analyze, making them difficult to utilize on a large scale. The current emphasis on reforms and accountability in the U.S. educational system creates a particularly pressing need to document and understand these initiatives at scale—which many have endeavored to do by drawing on survey data. Unfortunately, there are mixed results with respect to the validity of teacher surveys (Mayer, 1999). One notable area of challenge with respect to validity has been with questions that include terms associated with reforms (Burstein et al., 1995; Desimone & Le Floch, 2004), which is particularly problematic given that the need to assess reforms at scale often necessitates survey use.

In this study we describe how our analysis of teachers' interpretations of survey items, focused on mathematics education reforms, led us to reconceptualize the broader source of teachers' misinterpretations as an issue of fit between the researchers' intended interpretation and teachers' professional practice. We use the notion of teachers' professional practice to refer not only to what teachers do when providing instruction to groups of students in the classroom (i.e., their instructional practice) but also to the other things they do as a part of their job as teachers (e.g., lesson planning, interacting with colleagues; Chaiklin & Lave, 1993).

We came to our reconceptualization of teachers' "misinterpretations" of survey items through our analysis of a survey designed to understand particular learning opportunities within teachers' professional practice. Researchers have used surveys to understand a number of aspects of teachers' work, including their instructional practice, professional development, and with which colleagues they interact (i.e., teachers' social networks). Qualitative research suggests that the content of teachers' interactions dramatically influences their learning opportunities in the context of those interactions (Coburn & Russell, 2008; Horn & Little, 2010). Up to this point, information about the content of teachers' interactions has been collected with interviews and observations and focused on relatively small samples of teachers. We wanted to understand the content of teachers' advice-seeking interactions, with particular interest in the learning opportunities they afford, but we wanted to do so across a larger sample. Hence, our original goal was to develop a survey that would measure the content of mathematics teachers' advice-seeking interactions, at a level of specificity that would allow us to determine the quality of the learning opportunities those interactions held for teachers.

We designed a survey and conducted cognitive interviews to understand teachers' interpretations of the items. Through

this analysis we found that prior conceptualizations of sources of misinterpretation did not adequately account for all of the inconsistencies between the survey items and teachers' interpretations. We found it useful to reconceptualize the broader source of many of the misinterpretations as an issue of fit between the researchers' intended interpretation and teachers' professional practice. We argue that this reconceptualization has wider implications for the design of surveys and the interpretation of survey data. In this article, we describe how this reconceptualization emerged from our investigation of the following research question: What are the sources of inconsistency between teachers' interpretations and researchers' intended interpretations of survey items aimed at measuring the content of teachers' advice-seeking interactions?

Literature Review

As we examined teachers' interpretations of survey items aimed at measuring the content of their advice-seeking interactions, it was important to consider the broader contexts in which they worked, the nature of the aspects of professional practice we were trying to measure with surveys, and common sources of misinterpretation of surveys of teachers' professional practice.

The Context of Mathematics Instruction

Over the past several decades, there has been increasing emphasis on test-based accountability in schools in the United States, solidified by the No Child Left Behind Act of 2001 (No Child Left Behind [NCLB], 2002). NCLB required that all students achieve at the proficient level or higher (defined by each state) in mathematics and language arts on state standardized tests by 2014, with schools being held accountable for all groups of students making adequate yearly progress. The emphasis on students' achievement on state standardized tests under NCLB led some districts and school principals to emphasize increasing test scores as their primary goal (O'Day, 2004).

In addition to the accountability climate fostered by NCLB legislation, new reform goals and standards for students' mathematical learning have been put in place over the past two decades (e.g., see National Council of Teachers of Mathematics [NCTM], 1989, 2000; National Governors Association for Best Practices & Council of Chief State School Officers, 2010). These goals for students' mathematical learning imply new expectations for mathematics teachers' instructional practice. The Curriculum and Evaluation Standards and Principles and Standards for School Mathematics documents published by the NCTM (1989, 2000) and the more recent Common Core State Standards (National Governors Association for Best Practices & Council of Chief State School Officers, 2010) reflect a consensus within the mathematics education research and policy communities for comprehensive reforms to traditional mathematics teaching and learning. Changes to mathematics instruction include the use of more open-ended, challenging tasks and opportunities for students to work with classmates on these tasks and discuss multiple solution strategies (Stein, Engle, Smith, & Hughes, 2008; Stein, Grover, & Henningsen, 1996). This type of instruction has been characterized as "ambitious" because it is challenging for teachers and has ambitious goals for students (Lampert & Graziani, 2009).

With these new aspects of instruction come increasing demands for teachers with respect to their knowledge of mathematics, understanding of student thinking, and pedagogical skills (Charalambous, 2010; Hiebert et al., 1997; Hill et al., 2008; Stein et al., 2008). Therefore, developing these types of instructional practices requires considerable learning on the part of teachers (Ball & Cohen, 1999; Kazemi & Franke, 2004; Stein, Smith, & Silver, 1999; Thompson & Zeuli, 1999). Through interactions with students or colleagues, planning lessons, and participating in professional development, teachers learn on the job (e.g., Franke & Kazemi, 2001; Remillard, 2000; Sun, Wilhelm, Larson, & Frank, 2014). In addition to functioning as a resource for teachers' learning, teachers' interactions with colleagues play an important role in shaping the local interpretation and implementation of reforms (e.g., Coburn, 2001; Coburn, Russell, Kaufman, & Stein, 2012; Penuel, Riel, Krause, & Frank, 2009). As such, there is great value in understanding teachers' interactions with colleagues, both as a resource for teachers' professional learning and as a site for the negotiation of meaning of reforms.

Teachers' Advice-Seeking Interactions

Teachers' social networks have been increasingly documented and analyzed to understand teachers' professional learning and the spread of educational innovations and reforms (e.g., Coburn & Russell, 2008; Coburn et al., 2012; Frank, Zhao, & Borman, 2004; Moolenaar, 2012; Penuel et al., 2009; Sun et al., 2014). Although studies have shown that interacting with colleagues can support learning and the improvement of practice, learning opportunities through advice-seeking interactions can vary depending on what is actually discussed. Close analysis reveals stark differences in the potential of these conversations to support teacher learning (e.g., Horn & Little, 2010). For example, if teachers talk abstractly about how their lessons went for the day without getting into any specifics of the lessons and reasons for success or failure, they are unlikely to learn much about instruction from that interaction. This is not to say that nothing will be learned from that interaction but that it is unlikely to deeply influence their teaching.

In their research on mathematics teachers' social networks, Coburn and Russell (2008) suggested that an important aspect of the content of teachers' interactions is the *depth* of those interactions. The depth of interaction is the degree to which an interaction offers opportunities to learn

or improve practice. We are specifically interested in the development of knowledge and practice that support ambitious mathematics instructional practice (i.e., instructional practice consistent with the reforms described above).

Coburn and Russell (2008) drew on observational and interview data to examine the depth of mathematics teacher interactions in the context of eight elementary schools engaged in reform aimed at developing ambitious instructional practices. Low-depth interaction was characterized by a focus on "surface structures and procedures, such as sharing materials, classroom organization, pacing, and how to use the curriculum" (Coburn & Russell, 2008, p. 212). In contrast, high-depth interaction "addressed underlying pedagogical principles of the approach, the nature of the mathematics, and how students learn" (Coburn & Russell, 2008, p. 212). In a related analysis, they found that engaging in high-depth interactions contributed to sustainability of reforms (Coburn et al., 2012). To our knowledge, only Coburn and her colleagues have considered the actual content of teachers' interactions within these networks. Surveys of teachers' social networks typically focus on structural aspects of these networks (e.g., presence of ties, direction of ties, and possibly frequency of interaction; Moolenaar, 2012). The use of observational and interview data allowed for more nuanced information about the potential learning opportunities within teachers' interactions, but this approach is time-consuming and expensive.

In our study, we were interested in making similar distinctions with regard to the depth of teachers' interactions, but the scope of our study (which involved 120 teachers across 30 schools in four districts) made it implausible to collect and analyze data following the approach of Coburn and colleagues. Instead, we set out to construct a survey that would enable us to measure the depth of interactions, or potential learning opportunities, of teachers' advice-seeking interactions.

Measuring Teachers' Professional Practice With Surveys

The current emphasis on reforms and accountability in the U.S. educational system creates a particularly pressing need for valid surveys in order to document and understand these initiatives at scale. Unfortunately, many studies have documented challenges associated with using surveys to measure teachers' professional practice. The majority of these studies have focused on teachers' instructional practice, and results have been mixed. In particular, researchers have established the validity of surveys to measure some aspects of instructional practice, such as individual indicators of specific instructional content (Burstein et al., 1995; Smithson & Porter, 1994) or composite measures of instructional strategies (Mayer, 1999). On the other hand, there has been less success developing valid surveys that measure some of the more nuanced aspects of teachers' instructional practice, particularly with regard to reform instructional strategies (Burstein et al., 1995) and the quality of their implementation (Mayer, 1999). In particular, Burstein et al. (1995) found that teachers' interpretations of terms associated with the reform mathematics movement were varied and inconsistent with their intent. For example, they found that teachers interpreted *problem solving* "in a narrower sense of solving traditional mathematics problems, rather than strategies for solving real-world or nonroutine problems" (p. 54).

In addition to being used to measure teachers' instructional practice, surveys have also been used as a way to measure teachers' professional learning opportunities more broadly (Desimone, 2009). Few studies have investigated the validity of surveys of professional learning opportunities. Desimone and Le Floch (2004) used cognitive interviews to investigate the validity of their survey of teachers' professional learning opportunities and found that they had difficulty validly measuring the form (e.g., formal professional development versus informal interactions) and content of teachers' professional learning opportunities. Desimone and Le Floch describe these dimensions as related to the professional development reform movement. In particular, survey items pertaining to informal interactions and the content of learning opportunities are more likely to involve terms associated with reforms. Therefore, as we set out to investigate sources of inconsistency in interpretation of a survey measure of another aspect of teachers' professional practice, it was particularly important to pay attention to terms associated with reform.

Theoretical Framing

Professional Practice

The notion of practice helps us conceptualize this work and relate it to other fields. Broadly conceived, nearly everything that humans do is considered a practice, both in their professional lives (e.g., ship navigating, blacksmithing, grading university exams) and their personal lives (e.g., shopping, interacting with friends) (Chaiklin & Lave, 1993). In this study we narrow our lens on practice to consider the day-to day-professional work, or professional practice, of mathematics teachers. As previously mentioned, we use *professional practice* to refer not only to instructional practice but also the other things they do as a part of their job as teachers.

Communities of Practice

The community of mathematics teachers defines the professional practice of teaching mathematics as they jointly engage in their work and interact with each other. Together, mathematics teachers constitute a community of practice (Wenger, 1998). In this analysis, we consider both the community of practice of mathematics teachers and the

community of practice of education researchers. We draw on this lens for conceptualizing meaning making and communication within and across these two groups. In particular, we argue that although the two communities have many points of compatibility in their goals for student learning, members of these two communities of practice experience vastly different day-to-day experiences around student learning and are held accountable for vastly different outcomes. For instance, mathematics education researchers are held accountable for publishing theoretically driven work that advances the field's knowledge about the teaching and learning of mathematics, whereas mathematics teachers tend to be held accountable for teaching in ways that result in satisfactory student outcomes on standardized tests (in addition to managing day-to-day interactions with students, colleagues, administrators, and parents).

This lens of communities of practice is not one that tends to be explicitly taken on in the literature on survey development. We argue in this article that this lens can be fruitful for making sense of difficulties in using surveys to measure aspects of teachers' professional practice.

A sociological construct that is commonly used in research that draws on a communities-of-practice lens is the notion of boundary objects, which Wenger (1998) defined as "objects that serve to coordinate the perspectives of various constituencies for some purpose" (p. 106). In this work, we conceive of our survey about teachers' professional advice-seeking interactions as a boundary object that coordinates the perspectives of mathematics education researchers and mathematics teachers as it serves its primary purpose of gathering valid information for mathematics education researchers. We were interested in how our survey of teachers' instructional advice-seeking interactions functioned in achieving its purpose for the mathematics education researcher community, by effectively communicating with the mathematics teacher community.

Method

Recall that our broad goal was to develop a survey to measure the content of teachers' instructional advice-seeking interactions; in this analysis we focus on sources of item misinterpretation. In the following paragraphs we describe the data sources and analysis methods.

Data Sources

The cognitive interview data presented here were gathered in the context of a larger research project that sought to address the question of what is needed to improve the quality of middle-grades mathematics teaching, and thus student achievement, at the scale of a large urban district (Cobb & Jackson, 2011; Cobb & Smith, 2008). The research team collaborated with the leaders of four large, urban districts that were attempting to achieve a vision of high-quality mathematics instruction that was compatible with the NCTM's (2000) Principles and Standards for School Mathematics. In each of the 4 years of the study (2007–2011), we collected several types of data to test and refine a set of hypotheses and conjectures about district and school organizational arrangements, social relations, and material resources that might support mathematics teachers' development of ambitious instructional practices at scale. One key focus of our data collection efforts was teachers' social networks. As described above, we were particularly interested in the depth of, or learning opportunities within, teachers' interactions.

In this study, we set out to understand how our participants interpreted a set of survey items pertaining to the content of their advice-seeking interactions. The items designed to measure the depth of teachers' interactions on the network survey were developed both inductively and deductively. Some of the survey items developed inductively from teachers' responses to an interview question about who they go to for advice or information about teaching mathematics and what types of things they talk about, which was asked of participants in the 1st year of the larger research project. Other survey items were developed out of current theories from the mathematics teacher learning and professional development literatures. The intent of the survey was to account for variation in learning opportunities arising from common interactional practices, by including six low-depth items (i.e., thought to have lower potential for teacher learning) along with six high-depth items, which were theoretically more likely to support teacher learning. We administered network surveys to all math teachers, math instructional coaches, and administrators in 30 middle schools across four large, urban districts in the 2nd through 4th years of the research project. Participants were asked to whom they had turned for advice about mathematics instruction during the previous year and what they talked about with that person (with the option of checking as many of the 12 items as they chose). The 12 items are shown in Figure 1.

Data for this study come from cognitive interviews conducted in the summer and fall of 2011, which followed the 4th year of data collection. Because many of the items were developed inductively from teacher interviews, we did not conduct cognitive interviews prior to using the survey items in Years 2 to 4 of data collection. Preliminary analysis of the survey data revealed a lack of patterns in the data, so we decided to conduct retrospective cognitive interviews in order to better understand how the items were being interpreted (Desimone & Le Floch, 2004; Karabenick et al., 2007). The goal of these interviews was to simulate how teachers make sense of the network survey while they are taking it by asking them to answer the questions on the network survey during the interview and think aloud as they decided whether or not they engaged in each type of interaction. Given that the cognitive interviews were conducted several months after participants had taken the survey, our intent was not to examine whether their answers were the

1. During this school year (including last summer), to whom have you turned for advice or information about teaching mathematics? Please write full first and last names (if known), and give a brief description of that person's role or position (e.g., teacher at my school, teacher at another school, assistant principal, principal, math coach, district math leader).
   Name: ☐
   Role: ☐

2. What type(s) of advice or information do you seek from this person? Please check all options that apply.
   - ☐ Doing mathematics problems together with discussions of different solution strategies
   - ☐ Discussing different ways students are likely to solve tasks
   - ☐ Discussing why some students didn't learn as expected in a lesson in order to plan for future instruction
   - ☐ Analyzing examples of student work in order to adjust instruction
   - ☐ Analyzing examples of student work to understand the different ways that students solve problems
   - ☐ Analyzing student work to see if students "got it"
   - ☐ Discussing how to make use of student solution strategies in whole class mathematical discussions
   - ☐ Discussing pacing
   - ☐ Discussing what materials to use for a lesson
   - ☐ After a lesson, sharing whether students "got it"
   - ☐ Sharing materials or activities
   - ☐ Updating one another on a student or students' progress in mathematics
   - ☐ Other (please specify):
     ☐

3. How often do you seek advice or information from this person?
   ○ Daily or almost daily ○ Once or twice per week ○ Once or twice per month ○ A few times per year

4. How influential is his/her advice on your work?
   ○ Not at all ○ Somewhat ○ Very

FIGURE 1. *Network survey.*

same, nor was it to ask them to explain the way in which they had previously responded to the survey. Instead, the interview would allow us to probe for examples of the kind of interactions they recalled as examples of the interactions they engaged in and determine the extent to which participants' interpretations aligned with our intent. In order to help interviewees understand the nature of the think-aloud protocol, we began the interview with a practice question about what kinds of TV shows they watched, first modeling the think-aloud process and then asking them to do the same.

We interviewed a sample of 12 teachers from just three of the four districts because teachers in one of the districts generally reported fewer opportunities to interact with colleagues about mathematics instruction. We developed a structured interview protocol to be used over the phone with surveys e-mailed to participants. The interviews lasted approximately 30 min and were audio recorded. We e-mailed participants two documents: the practice question about watching TV shows and the network questions. As described above, the practice question was utilized to help the participant understand the process of thinking aloud while responding to survey prompts. The network questions were the set of questions from the network survey (see Figure 1). As a part of the interview protocol, we asked participants to think aloud as they decided whether or not they would check each box. Once participants had answered the entire set of items pertaining to depth of interactions (see Question 2 in Figure 1), we followed up by asking for examples when they had not been provided as part of their response.

Analysis

In order to better understand the sources of interpretation inconsistency, we conducted our analysis in three phases: identifying consistent and inconsistent responses, classifying types of inconsistency in interpretation, and analyzing sources of inconsistency. In order to facilitate this analysis, all 12 of the cognitive interviews were transcribed with responses placed into a participant by item grid. This organizational approach facilitated analysis within and across participants and items to look for trends.

Phase 1: Interpretation inconsistency. Our first phase of analysis aimed to help us determine for which items our survey functions as a boundary object that facilitates effective communication between us, as mathematics education researchers, and our participating mathematics teachers. To this end, we coded participant responses for whether their interpretations were consistent with our intent. First, we generated a codebook to document the intended interpretation of each item. An example of the codebook for Item A is shown in Table 1. We then coded responses to each survey item

TABLE 1
Codebook for Item A

Number	Item	Intended interpretations show evidence that . . .
A	Doing mathematics problems together with discussions of different solution strategies	1. Participant describes talking with a colleague 2. Participant describes actually working through (i.e., solving) math problem(s) 3. There is evidence that more than one approach to a problem is discussed (this could include talk about correct and/or incorrect approaches)

according to whether the participant described an example that was consistent with our intended interpretation. Participant responses were coded as "yes" (the response is consistent with our intended interpretation), "no" (the response is not consistent with our intended interpretation), and "not coded" (we do not have enough information to determine if the response is consistent with our intended interpretation). The two authors independently coded responses as consistent, inconsistent, or not codable and then came to consensus on the consistency of each response.

Phase 2: Types of inconsistency in interpretation. In our second phase of analysis, we set out to classify types of inconsistency in interpretation so that we could later examine sources of inconsistency. This phase of analysis had two steps. We examined teachers' responses that were identified as inconsistent with the intended interpretation, inductively categorized the types of inconsistency that arose, and documented any other emergent themes (Strauss & Corbin, 1998). In doing so, four categories of types of inconsistency emerged: (a) "no talk with colleagues," which indicates that the participant did not describe an example that included conversation with colleagues; (b) "omit part," which indicates that the participant gave an example that matched the intended interpretation for part of the item but did not address part of the item; (c) "phrase-level inconsistency," which indicates that the participant inconsistently interpreted a particular word or phrase within the item; and (d) "item-level inconsistency," which typically involved omitting part of the item as well as inconsistently interpreting a word or phrase in the item.

Phase 3: Sources of inconsistency in interpretation. Before looking for emergent themes with respect to inconsistency in interpretation, we focused on conjectured sources of inconsistency. These included reform language, structural complexity, and depth of interactions. As described above, prior work on the validity of surveys of teachers' instructional practice suggested that terms associated with the reform mathematics movement have the potential to be problematic with respect to interpretation (Burstein et al., 1995). Therefore, we identified terms in survey items we thought might be interpreted by mathematics teacher participants in ways other than we, as members of the mathematics education researcher community of practice, intended (e.g., solutions strategies, tasks, whole-class mathematical discussions).

Our conjecture that items with a linguistically complex structure would contribute to inconsistency is also grounded in the literature. The tension between reducing vagueness and minimizing structural complexity is well documented in the survey design literature (Fowler & Consenza, 2008; Groves et al., 2004). Many of the terms we use in everyday language are inherently vague, and surveys may aim to be more concrete and specific. For example, in the context of measuring teacher interactions, if the item asks respondents if they talk about "analyzing student work," then we do not know if they are doing this in ways that we would deem as "high-depth" or if they are analyzing student work in more superficial ways. To resolve this ambiguity, the survey designers added modifiers to a number of items to clarify the purpose. However, these items then became more complex. When coding for complexity of items a priori, we classified items that have two or more separate clauses as complex.

In addition, the "high-depth" items tended to originate from the teacher-learning literature, which often highlights rare but productive forms of professional practice that are more closely aligned to with the concerns of mathematics education researcher community of practice. In contrast, "low-depth" items tended to originate from interviews with participants that took place in the 1st year of the larger project, prior to the development of the survey instrument; thus the practices corresponding to low-depth items are likely to be more frequently aligned with the practices of members of the mathematics teacher community. As such, high-depth items might have been more likely to be interpreted inconsistently than low-depth items. We classified each of the survey items as high- or low-depth based on the teacher-learning literature.

After examining the ways in which our a priori conjectures regarding reform language, structural complexity, and depth accounted for inconsistency in interpretation, we looked for emergent themes in the data, with special attention to teachers' professional practice, to explain the inconsistencies in interpretation that could not be accounted for with the a priori conjectures.

Findings

The three phases of analysis built on one another to uncover sources of inconsistency between teachers' interpretations and researchers intended interpretations. We begin

TABLE 2
Interpretation by Item With Coding Results

				Interpreted as intended?				Type of inconsistency			
Depth	Complex	Item	Wording	Yes	No	% No	Not coded	No talk with colleague	Omit part	Phrase	Item
High	Y	**A**	**Doing mathematics problems together with discussions of different *solution strategies***	6	**3**	**33.3**	3	0	0	2	1
High	N	B	Discussing different *ways students are likely to solve tasks*	8	2	20.0	2	0	0	2	0
High	Y	**C**	**Discussing why some students didn't learn as expected in a lesson in order to plan for future instruction**	1	**9**	**90.0**	2	0	3	0	6
High	Y	**D**	**Analyzing examples of *student work* in order to adjust instruction**	4	**7**	**64.6**	1	0	3	1	3
High	Y	E	Analyzing examples of *student work* to understand the *different ways that students solve problems*	5	1	16.7	6	0	1	0	0
Low	Y	F	Analyzing *student work* to see if students "got it"	8	1	11.1	3	0	1	0	0
High	Y	**G**	**Discussing how to make use of *student solution strategies* in *whole-class mathematical discussions***	4	**4**	**50.0**	4	1	1	2	0
Low	N	H	Discussing pacing	9	1	10.0	2	1	0	0	0
Low	N	I	Discussing what materials to use for a lesson	7	3	30.0	2	0	3	0	0
Low	N	J	After a lesson, sharing whether students "got it"	3	0	0	9	0	0	0	0
Low	N	K	Sharing materials or activities	10	0	0	2	0	0	0	0
Low	N	L	Updating one another on a student or students' progress in mathematics	8	1	11.1	3	1	0	0	0

Note. Italics indicate reform language.

with a summary of consistency of interpretation and variation by item and then describe the emergent findings with respect to types of inconsistency and sources of inconsistency.

Results of the first two phases of analysis are summarized in Table 2. Overall, we found that the mean rate at which items were interpreted as intended was about 72%. The overall rate at which each item is not interpreted as intended is represented in the "% No" column for each item as shown in Table 2. There is considerable variation in the rate of inconsistent interpretation by item, with rates ranging from 0% to 90%. In bold are the four most problematic items with rates of inconsistent interpretation over 30%. In the table we indicate how each item was coded for a priori categories of structural complexity and depth as well as the frequency with which each code was as a type of inconsistency (i.e., no talk with colleague, omit part, phrase-level inconsistency, and item-level inconsistency). Although some inconsistencies could be explained by our a priori conjectures about problematic item features, the variation in whether the conjectured item features were problematic along with several unanticipated interpretations led us to a more general finding: Inconsistency in interpretation was often due to the fit between the researchers' intended interpretation for the item and the teacher's professional practice. In the subsequent sections, we summarize our findings with respect to the a priori conjectures about item features and then expand on our findings that arose from emergent coding of inconsistencies.

A Priori Conjectures About Item Attributes

There is evidence that the item attributes of reform language, structural complexity, and depth contributed to inconsistency of interpretation, as shown in Table 2. With regard to reform language, we identified *solution strategies* and *whole-class mathematical discussions* as potentially problematic phrases, which proved to be the case. We did not anticipate that the phrase *student work* would be a source of misinterpretation, and we found that it was interpreted inconsistently when used in some items but not others. For instance, participants' interpretations of *student work* contributed to inconsistent interpretations of Item D, but *student*

work was also used in Items E and F, which were interpreted inconsistently less than 20% of the time. In particular, *student work* was interpreted inconsistently when the purpose of its analysis was to inform instruction, which could suggest that this is not the purpose for which teachers predominantly use student work (e.g., using student work to do things like assign grades or to identify which students need additional supports outside of class)—pointing to an issue with the item's fit with teachers' professional practice rather than with the phrase itself.

Overall, six of the 12 items used terms we classified as including reform language (indicated in italics within the item), but only four of 12 items (A, C, D, G) were inconsistently interpreted more than 20% of the time. Three of the four most problematic items used reform language; only Item C did not. Interestingly, Item C was the most inconsistently interpreted item, and like Item D, it makes reference to informing instruction (in this case, planning for future instruction). As such, we conjecture that, in teachers' practice, these kinds of activities (i.e., analyzing student work or discussing why students did not learn as expected in a lesson) are often done for purposes other than informing instruction (e.g., identifying students for additional test preparation tutoring), and therefore the intended interpretation of the item did not fit with teachers' practice.

There is also evidence that structural complexity and depth contributed to interpretation inconsistency; six of 12 items were structurally complex, and six of 12 items were high depth (although not the exact same six as are structurally complex; see Table 2). This may suggest that complexity was often needed to describe high-depth interactions without ambiguity. All four items that were misinterpreted more than 20% of the time were both complex and of high depth. Taken together, there is significant overlap between complex items and high-depth items; there is only one item that is high depth that is not structurally complex (B) and only one item that is complex that is not high depth (F). Interestingly, both of these items (B and F) were typically interpreted as intended.

Reform language heavily overlapped with structurally complex and high-depth items, whereas only one of the low-depth items involved reform language. The only high-depth item that did not have reform language was Item C, but this item was structurally complex, as previously mentioned, and had, in fact, the highest rate of inconsistency of interpretation. Item C was particularly interesting because it had the highest rate of item-level inconsistency of interpretation, indicating a more general type of inconsistency; this result is discussed in greater detail in the next section.

Types of Interpretation Inconsistencies

The way in which interpretation inconsistencies are distributed across our four types is shown in Figure 2, with the greatest number of inconsistencies arising from omission of part of the item (38% of inconsistencies). In the following paragraphs we elaborate on the three most common types of inconsistency—omit part, phrase level, and item level—in order of frequency.

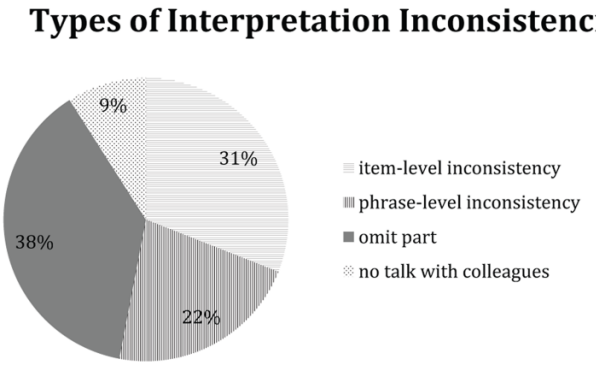

FIGURE 2. *Distribution of interpretation inconsistencies.*

It is plausible that inconsistent interpretations that arise from the omission of part of an item could arise as a result of structural complexity (e.g., teachers ignored or just forgot to mention that aspect of the item), or teachers might choose to ignore part of an item that fails to align with their professional practice. Recall that items were classified as complex if they included a modifier in addition to the main clause. For example, Item C, "Discussing why some students didn't learn as expected in a lesson in order to plan for future instruction," includes the clause "discussing why some students didn't learn as expected in a lesson" and also the modifier (here with the purpose of the activity described in the modifier) "in order to plan for future instruction." One teacher who omitted the part of the item about planning for future instruction gave the following example: "[The curriculum] seems to be moving much slower than last year and we were discussing why they are not learning like we expected them to learn, like they did last year. And what the differences for that might be." In this case, the teacher describes a conversation about why students did not learn as expected but does not address planning for future instruction. She may have omitted the planning part of the item because she forgot about it due to the item's complexity, or she may have omitted the planning part of the item because she does not simultaneously discuss why students are not learning as expected while planning for future instruction. The source of interpretation inconsistency could be structural complexity or the fit of the item with the teachers' professional practice. This kind of inconsistency (i.e., omitting part of the item) constituted 38% of the inconsistencies in interpretation identified in our analysis.

Phrase-level inconsistencies in interpretation (including those arising from reform language), which are well documented in the research literature, account for 22% of the

inconsistencies identified in our analysis (i.e., a total of seven instances in the phrase-level-inconsistency column of Table 2). Phrase-level inconsistencies have the potential to arise from situations in which teachers do not know or understand the meaning of a particular word or phrase or from situations in which the meaning the word or phrase holds for teachers differs from researchers' intent in the context of teachers' current professional practice.

For example, the phrase *whole-class mathematical discussions* contributed to two of the four inconsistent interpretations of Item G. Those two teacher participants described these discussions in ways that included any occasion when the whole class was listening to a single speaker (e.g., during an introductory lecture or guided practice). We intended it to refer to whole-class discussions as conversations in which multiple participants contribute ideas and that take place following students' work on a task. These two inconsistent interpretations provide another example of how our intended interpretation did not fit with some teachers' instructional practice. Overall, participants' interpretations of particular phrases were not as problematic as we anticipated, but there were some phrases that were interpreted inconsistently, which contributed to overall inconsistency in interpretation of items. In general, phrase-level inconsistency in interpretation was indicative of a lack of fit between the item and teachers' professional practice.

The last type of inconsistency identified in our analysis was item-level inconsistency in interpretation and occurred for only a few items. This category does not map on directly to structural complexity, reform language, or depth. Often the teacher's response had some overlap with the intended interpretation but was missing the essence of the item in ways that suggest a misfit with teachers' professional practice. For example, in Item C, "Discussing why some students didn't learn as expected in a lesson in order to plan for future instruction," the majority of the participants who did not interpret the item as intended did not describe an example in which they were discussing *why* some students did not learn as expected. One teacher's example is typical of the examples given by teachers with item-level misinterpretations:

> We just look at the way I presented it and then she [the person from whom advice was being sought] would help me with another way to present it to the students the second time around, if I realized that the students didn't get it the first time.

This overlaps with aspects of the intended interpretation of the item (e.g., talking about prior instruction and using that to plan for future instruction), but the issue is that the essence of the item was to actually talk about *why* rather than just talking about the fact that students did not learn as expected. Examples of reasons why could include some aspect of the way the teacher had taught a lesson or why the lesson failed to connect with or build on students' prior knowledge. Given the responses from participants, it seems likely that they might not actually talk about why students did not learn as expected at that level of detail and might instead stick to talking about the fact that they need to "reteach" the content. From our perspective, understanding why students did not learn as expected is fundamental to making instructional adjustments that will be effective. This is an indication that perhaps the item itself was problematic because few teachers actually talk with their colleagues at this level of depth, but they were able to describe conversations that involved many of the features of the item. In other words, the inconsistency in interpretation of the item arose from the fact that the described interaction did not fit with most teachers' professional practice.

A Central Theme in Participant Responses: The Testing Context

As we looked across our cognitive interview data, we noted that testing and use of test data was an unexpected but frequent theme in teachers' responses. This was interesting to us, as none of the items explicitly asked about testing or data use. However, given the increasing emphasis on testing and use of data to inform instruction in schools in the U.S. public education system, it is perhaps unsurprising that teachers' responses were often framed in this way. The prevalence of testing in the examples teachers provided further substantiated the need for attention to teachers' professional practice.

We noted that 10 of the 12 teachers interviewed reported turning to colleagues about issues related to district or state testing (including both test preparation and use of test data) in their cognitive interviews, spread across eight different items. Most references to use of testing data involved making decisions about pacing (when to move on), what to reteach, and which students needed tutoring or other interventions.

From our perspective, this prevalence of testing and use of test data in teachers' descriptions of particular types of interactions highlights the differences between the mathematics education research community of practice and the mathematics teacher community of practice as we interacted around the survey. Although we were very aware of the accountability emphasis in their schools, we did not anticipate that they would find testing-related examples consistent with our survey items. We take this as further evidence that teachers interpreted the survey items in ways that fit with their professional practice.

In summary, there is evidence that our a priori conjectures about reform language, structural complexity, and depth contributed to the inconsistency in interpretation of items. However, the emergent category of testing-related interpretation, the fact that some phrase-level inconsistencies in interpretation were specific to items rather than particular terms, and the presence of item-level inconsistencies

suggest that a broader source of inconsistency in interpretation arises from the fit of the phrase or item with teachers' professional practice. Further, fit between items and teachers' professional practice contributed to all of the types of inconsistencies (i.e., omit part, phrase level, item level, and no talk with colleagues) we found. Below we elaborate on these findings and discuss their implications. First, we briefly describe a few limitations of this analysis.

Limitations

In this analysis we were attempting to understand how participants in the larger study interpreted survey items. We selected a sample of 12 participants who varied in district membership, years of experience teaching, and instructional quality. It is possible that our sample was not representative of the larger population in terms of how the participants interpreted the survey items, but we have no reason to believe that that was the case. We believe that our sampling limitations are unlikely to dramatically affect our findings.

Another limitation of this analysis is the potential for social desirability to influence how participants responded in cognitive interviews. Although we clearly stated within the cognitive interviews that there were not any right answers and we were just interested in understanding how they interpreted the items on the survey, it is possible that participants might have tried to answer "yes" to as many of the options as possible, which could have contributed to our findings that for some items, teachers interpreted the items in ways that fit with their professional practice but were outside of our intended interpretation.

Discussion

Given the current emphasis and the general importance of assessing innovations at scale and the widespread use of surveys for this purpose, understanding threats to validity of survey data is particularly important. In this article we set out to describe how our investigation of sources of inconsistency between teachers' interpretations and researchers' intended interpretations of survey items aimed at measuring the content of teachers' advice-seeking interactions led us to a reconceptualization of teachers' misinterpretations of survey items as an issue of fit between the researchers' intended interpretation and teachers' professional practice. We found this reconceptualization helpful for making sense of both anticipated sources (e.g., reform language, structural complexity) and unanticipated sources of inconsistency in interpretation. In the following paragraphs, we first discuss how our findings fit with prior research, then discuss questions that arise from our findings, and finally, discuss implications for the design of surveys.

We had several a priori conjectures about sources of inconsistency in interpretation: structural complexity, reform language, and depth of interaction. First, recall that a central challenge with respect to items' structural complexity is one of trying to decrease ambiguity while minimizing complexity (Fowler & Consenza, 2008; Groves et al., 2004). In our case, adding modifiers often served to make more general phrases, like *discussing student work*, less ambiguous so that we could understand how and why teachers discuss student work, because the methods and purposes offer very different potential for teachers' professional learning. We found that teachers did omit the modifiers on many occasions, but it is unclear whether they omitted modifiers just because the items were too complex or if they were also omitted because the modifiers narrowed the description of the interaction to one that did not fit with their professional practice. Further studies should investigate this with cognitive interviews with more detailed follow-up probes to ask about the reasons for teachers' interpretations.

A number of studies have suggested that reform language is problematic for teacher surveys (Burstein et al., 1995; Desimone & Le Floch, 2004; Mayer, 1999). Recall, for example, that Burstein et al. (1995) found that teachers misinterpreted *problem solving.* We similarly found that teachers' interpretations of phrases associated with the mathematics reform movement did not match with our intended interpretation. We argue that their findings and ours, although they do pertain to reform language, are indicative of a larger issue of the fit of reforms with teachers' professional practice.

Our investigation of the high-depth items, the items that described interactions with greater potential for teacher learning, further revealed the importance of fit with teachers' professional practice. Although high-depth items tended to be marked by use of reform language and structural complexity, we found that the high-depth items were also more likely to be interpreted by teachers in ways that were inconsistent with our intent. The items that were developed out of the mathematics teacher-learning literature are notably rare, and although teachers sometimes understood the phrases that were used, they described an example that did not fit with the intended essence of the practice. We see this as indicative of a mismatch between researchers' intended interpretation and teachers' professional practice.

Last, the emergence of the data-and-testing theme was another indicator of the importance of considering teachers' professional practice. Although none of our items specifically mentioned data or testing, the majority of the participants brought that frame to their interpretations of one or more items. In many cases, they described examples that fit with our intended interpretations but described something in the context of testing or data, and in a few cases, they described an example that was outside of our intended interpretations. In either case, it was important to consider the alignment between our participating teachers' current professional practice and the intended interpretation of the item.

Our findings with respect to inconsistency of interpretation due to fit with teachers' professional practice lead to several questions about surveying teachers about their practice. First, is it the case that we can validly survey teachers only about practices that are part of their *current* professional practice? If this is the case, this would imply that their understanding of the terms implies that they engage in the practice, and yes-or-no questions about whether they engage in the practice would then potentially be redundant. Our data offer some hope that we can validly ask teachers about some practices that are not included in their current professional practice. For instance, in response to Item E, "Analyzing examples of student work to understand the different ways that students solve the problem," one teacher responded,

> I didn't do that as much, that's a good idea, maybe I should. I didn't. So, I would say no for that one. It is something to look into in the future. The whole analyzing student work and things like that was challenging. . . . And a lot of stuff there wasn't multiple ways to do it, so that's why, yeah.

On the basis of this response, we concluded that this teacher interpreted the item as intended even though he does not actually do this with his colleagues. This provides some hope that we, as education researchers, can ask teachers about practices that are outside of their current professional practice and teachers can still interpret them as intended. But when is this possible? Why did this teacher interpret this item as intended, but (mis)interpret other items in ways that suggest those kinds of interactions are outside of his current professional practice?

Further, if we are interested in reforms or innovations that seek to change teachers' professional practice, how do researchers design surveys to measure such changes when researchers are likely to ask about things that are outside of teachers' current practice? Is there a zone of proximal development for teacher development that not only influences their learning but also affects researchers' ability to validly survey teachers' development? There is still much to learn about how to validly assess teachers' practice using surveys. Future research should take up these questions as the need to accurately and efficiently measure teachers' practice at scale is not likely to diminish.

In this spirit, we describe some approaches to survey development that have the potential to account for some of our issues with interpretation. First, reducing the complexity of the items themselves, by using a tiered approach, has the potential to help to combat some issues with interpretation. For example, we might first ask an initial question about a general idea (e.g., "Do you analyze student work together?") and then ask more specifically about the purpose (e.g., "in order to adjust instruction"). Second, with respect to issues of fit with teachers' professional practice, perhaps using representations would give respondents an image of the practice they are being asked about. Such representations might include vignettes (i.e., short descriptive episodes; e.g., Kaufman et al., 2014), short video clips (e.g., Kersting, Givvin, Sotelo, & Stigler, 2010), or cartoon images (e.g., Herbst & Kosko, 2014). In the context of network surveys that aim to measure the content of interactions, participants might select those representations that best resemble their interactions with particular colleagues. Under this approach, respondents get to "observe" an interaction and decide if it is similar to how they interact with colleagues, rather than interpreting a general description of a type of interaction and determining whether they think their interactions are sufficiently aligned with that description to say that they engage in that kind of conversation with a particular colleague. Future work should continue to investigate this approach of using representations to measure different aspects of teachers' professional practice.

Conclusion

Although we offer some possibilities for other survey designs, our primary intent is to suggest the utility of a different lens for thinking about survey design and item misinterpretation. By drawing on the communities-of-practice lens, we reconceptualized issues of misinterpretation of survey items as instances of inconsistency in interpretation of a boundary object between the mathematics education researcher community of practice and the mathematics teacher community of practice. Although our survey of teachers' advice-seeking interactions helped us reconceptualize survey misinterpretation, we believe that bringing a communities-of-practice lens to survey development and interpretation is useful for thinking about reliability and validity of other teacher surveys and other surveys of education as well (see also Hill, 2005). The oft-cited gap Jordan (1989) documented between the language used to talk about professional work, or practice, and the reality of doing that work is reason to believe that inconsistency in interpretation between different communities of practice also extends beyond education.

Therefore, as we continue to use surveys as a means to efficiently assess educational innovation at scale, it will be important to consider the different communities of practice involved and how the surveys represent boundary objects with negotiated meanings. Cognitive interviews are an important first step in understanding survey item interpretation. Perhaps even when using previously validated surveys, it is important to conduct cognitive interviews to check that the survey is valid with a new population (i.e., is interpreted similarly by a different community of practice).

Acknowledgments

The authors were supported by the Institute of Education Sciences (IES) predoctoral research training program, Grant Number R305B080025, and the IES postdoctoral research training program,

Grant Number R305B080008, respectively. The data were collected in the context of a larger study, supported by the National Science Foundation under Grant Numbers ESI-0554535 and DRL-0830029. The opinions expressed do not necessarily reflect the views of the U.S. Department of Education or the National Science Foundation. We thank members of the research team for the larger project from which these data come, including Paul Cobb, Tom Smith, Kara Jackson, Erin Henrick, Dan Berebitsky, Kenneth Frank, Lynsey Gibbons, Charlotte Dunlap, and Adrian Larbi-Cherif.

References

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession* (pp. 3–32). San Francisco, CA: Jossey-Bass.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T. H., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Chaiklin, S., & Lave, J. (1993). *Understanding practice: Perspectives on activity and context*. Cambridge, UK: Cambridge University Press.

Charalambous, C. Y. (2010). Mathematical knowledge for teaching and task unfolding: An exploratory study. *Elementary School Journal*, *110*(3), 247–278. doi:10.1086/648978

Cobb, P., & Jackson, K. J. (2011). Towards an empricially grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development*, *13*(1), 6–33.

Cobb, P., & Smith, T. M. (2008). District development as a means of improving mathematics teaching and learning at scale. In K. Krainer & T. Wood (Eds.), *Participants in mathematics teacher education: Individuals, teams, communities, and networks* (Vol. 3, pp. 231–254). Rotterdam, Netherlands: Sense.

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145–170.

Coburn, C. E., & Russell, J. L. (2008). District policy and teacher's social networks. *Educational Evaluation and Policy Analysis*, *30*(3), 203–235.

Coburn, C. E., Russell, J. L., Kaufman, J. H., & Stein, M. K. (2012). Supporting sustainability: Teachers' advice networks and ambitious instructional reform. *American Journal of Education*, *119*(1), 137–182. doi:10.1086/667699

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199.

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.

Fowler, F. J., & Consenza, C. (2008). Writing effective questions. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 136–160). New York, NY: Lawrence Erlbaum Associates.

Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education*, *77*, 148–171. doi:10.1177/003804070407700203

Franke, M. L., & Kazemi, E. (2001). Teaching as learning within a community of practice. In T. Wood, B. S. Nelson, & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 47–74). Mahwah, NJ: Lawrence Erlbaum.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Herbst, P., & Kosko, K. W. (2014). Using representations of practice to elicit mathematics teachers' tacit knowledge of practice: a comparison of responses to animations and videos. *Journal of Mathematics Teacher Education*, *17*(6), 515–537.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., Murray, H., . . . Human, P. (1997). *Making sense: Teaching and learning mathematics with understanding*. Portsmouth, NH: Heinemann.

Hill, H. C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, *19*(3), 447–475.

Hill, H. C., Blunk, M. L., Charalambos, C. Y., Lewis, J. M., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511. doi:10.1080/07370000802177235

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal*, *47*(1), 181–217.

Jordan, B. (1989). Cosmopolitical obstetrics: Some insights from the training of traditional midwives. *Social Science & Medicine*, *28*(9), 925–937.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., . . . Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151.

Kaufman, J. H., Engberg, J., Hamilton, L. S., Hill, H. C., Umland, K., Yuan, K., & McCaffrey, D. F. (2014, April). *Anchoring measures of teacher instruction*. Paper presented at the annual conference of the American Educational Research Association, Philadelphia, PA.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. *Journal of Mathematics Teacher Education*, *7*, 203–235. doi:10.1023/B:JMTE.0000033084.26326.19

Kersting, N., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, *61*(1/2), 172–181.

Lampert, M., & Graziani, F. (2009). Instructional activities as a tool for teachers' and teacher educators' learning in and for practice. *Elementary School Journal*, *109*(5), 491–509. doi:10.1086/596998

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–45.

Moolenaar, N. M. (2012). A social network perspective on teacher collaboration in schools: Theory, methodology, and applications. *American Journal of Education*, *119*(1), 7–39. doi:10.1086/667715

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for mathematics*. Washington, DC: Author.

No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. (2002).

O'Day, J. A. (2004). Complexity, accountability, and school improvement. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 15–43). New York, NY: Teachers College Press.

Penuel, W. R., Riel, M., Krause, A. E., & Frank, K. A. (2009). Analyzing teachers' professional interactions in a school as social capital: A social network approach. *Teachers College Record*, *111*(1), 124–163.

Remillard, J. T. (2000). Can curriculum materials support teachers' learning? Two fourth-grade teachers' use of a new mathematics text. *Elementary School Journal*, *100*(4), 331–350.

Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum. The Reform Up Close Study*. New Brunswick, NJ: Consortium for Policy Research in Education.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, *10*(4), 313–340. doi:10.1080/10986060802229675

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, *33*(2), 455–488. doi:10.3102/00028312033002455

Stein, M. K., Smith, M. S., & Silver, E. A. (1999). The development of professional developers: Learning to assist teachers in new settings in new ways. *Harvard Educational Review*, *69*(3), 237–269.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research*. Thousand Oaks: Sage.

Sun, M., Wilhelm, A. G., Larson, C., & Frank, K. A. (2014). Exploring colleagues' professional influence on mathematics teachers' learning. *Teachers College Record*, *116*(6).

Thompson, C. L., & Zeuli, J., S. (1999). The frame and the tapestry: Standards-based reform and professional development. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 341–375). San Francisco, CA: Jossey-Bass.

Wenger, E. (1998). *Communities of practice*. Cambridge, UK: Cambridge University Press.

Authors

ANNE GARRISON WILHELM is an assistant professor of mathematics education at Southern Methodist University. Her research interests are the enactment of ambitious mathematics instruction, mathematics teachers' on-the-job learning, and the complexities of measuring teaching and learning at scale.

CHRISTINE ANDREWS-LARSON is an assistant professor of mathematics education at Florida State University. She is interested in student reasoning in secondary and undergraduate mathematics, teachers' pedagogical reasoning, and how teachers learn to make sense of and leverage students' mathematical reasoning in equitable ways.