Content across communities:
Validating measures of elementary mathematics instruction

Abstract

In recent years, scholars have problematized terms used to describe instruction on teacher survey instruments. When scholars, observers, and teachers employed terms like "discuss" and "investigate," these authors found, they often meant to describe quite different events (Mayer 1999; Spillane & Zeuli 1999; Stigler, Gonzales, et al, 1999). This paper problematizes another set of terms often found on survey instruments, those describing mathematical content. To do so, it examines terms such as "geometry," "number patterns" and "ordering fractions" for rates of agreement and disagreement between teachers and observers participating in a field pilot of an elementary mathematics daily log. Using interviews, written observations, and reflections on disagreements, this paper is also able to ask why disagreements occurred. Sources of disagreement included problems with instrument design, memory/perception, and, notably, differences in the way language is used in different communities – university mathematicians, elementary teachers, and mathematics educators – to give meaning to subject matter terms. Theoretical and practical implications of these sources of disagreement are explored.

In recent years, scholars have used large-scale survey techniques to gauge the effectiveness of reform efforts, professional development initiatives, and other interventions into the teaching and learning process (e.g. Colleague & Author, 2000,2001; Garet, Birman, et al 2001; Mayer 1999; Porter, Floden, et al 1996; Spillane & Zeuli, 1999; Stecher & Chun, 2000; Supovitz & Turner, 2000). Compared with past efforts to evaluate such programs and initiatives, survey instruments offer many benefits: the ability to gather data on a large number of lessons; the possibility of providing accurate population estimates for particular activities, and exploring relationships among policy, school environments, and teachers' practice; and the capacity to focus on instruction as reported by teachers. However, survey techniques also suffer from concerns about validity -- that is, the question of whether teachers' reports on survey instruments accurately represent their actual classroom practice.

One problem with the use of surveys is that they rely on words and phrases to describe instruction -- yet in the U.S., at least, language for instruction is underdeveloped and imprecise, and definitions can vary across communities which use those words or phrases. For instance, scholars have documented variations in meaning when words like "investigate" or "discuss" are used to represent particular methods of classroom work (Mayer, 1999; Spillane and Zeuli, 1999; Stigler, Gonzales, et al, 1999). Scholars have less frequently examined the validity of teachers' interpretations of terms associated with subject matter itself – e.g., in mathematics, "proof," "patterns," or "procedures." Because of the importance of such terms within educational research today, and

because the possibility exists that instruments measuring teachers' content coverage might be improved, this paper hopes to clarify why validity problems arise in the use of content terms on survey instruments.

To do so, this paper presents an analysis of data produced during a pilot of a daily mathematics log, a survey instrument which gave elementary school teachers and observers the opportunity to report on both broad content coverage (e.g., basic facts, geometry, fractions, algebraic reasoning) and finer content breakdowns for a limited number of topics. Some of the log's terms (e.g., geometry, algebra) are used by university mathematicians to demarcate particular fields of study and analysis; others (e.g., inequalities, ordering fractions) are mathematical terms, but used most often by teachers and others to specifically describe elementary school mathematics; still others (e.g., problem solving, number patterns) have been introduced or redefined by mathematics educators seeking to improve school mathematics. By analyzing how instrument writers, teachers, and classroom observers used the log during this field trial, we argue that terms describing mathematical content cannot be taken for granted as an agreed-upon lexicon in U.S. classrooms. We explore reasons for disjunctures, then comment on theoretical and practical implications of this problem.

.

Measuring Instructional Content

Most studies designed to validate teacher survey instruments, including daily logs, have focused on teachers' use of specific instructional practices — for example, teachers' use

of small group work, manipulatives, or discussion. These studies have reported mixed results. Mayer (1999) found that observed and self-reported composite teacher scores on measures of reform mathematics teaching correlated highly, at .85. While teachers inflated their self-reports of NCTM-aligned approaches to mathematics teaching, they did so systematically, maintaining their position relative to one another on the scales. Burstein, McDonnell, and other researchers at RAND (1995) found that survey data on instructional processes can provide an accurate picture of teachers' classroom practices. However, they also argued that this accuracy derived from the conventional and stable nature of teachers' practice, for the teachers whom they studied relied extensively on lecture and homework review, and reported little variation in their practices. Studies by Stigler, Gonzales, and others (1999) and Spillane and Zeuli (1999), although not technically validation studies, found that although some teachers reported extensive use of NCTM-aligned instructional approaches, few actually appeared to do so in their classroom.

Less is known about the validity of teachers' reports of subject matter content coverage. Smithson and Porter (1994), writing to describe validation work done in conjunction with the Content Determinants study (Porter, Floden et al 1986), examined the level of agreement between observers and teachers over 62 high school mathematics and science lessons. They found that for all gross content breakdowns (i.e., averaging intercoder agreement for algebra, geometry, probability), agreement between teachers and observers ranged between .61 and .80, depending upon the

method used to construct the measure. For finer content distinctions (i.e., averaging scores for all subtopics, such as variable, expressions, linear equations or inequalities), agreement ranged between .49 and .70, again depending upon the calculation method selected. These authors, however, did not identify patterns in teacher-observer agreement across subject matter; nor did they hypothesize why such disagreements might occur.

Another effort to validate indicators of subject matter content was included in the work of Burstein, McDonnell, and other researchers at RAND (1995). By comparing 70 secondary school mathematics teachers' logs of daily practice to artifacts from their classroom practice – textbooks, daily assignments, exams and quizzes – RAND researchers determined that survey reports reasonably accurately portray whether a topic has been taught (p. 29). The researchers noted that while fine-grained reports of the time spent on particular topics were not particularly reliable, more general estimates of time-on-task were: agreement between teachers' reports and classroom artifacts ranged from 42 to 71 percent, given 1-point leeway within a 5-point scale.

The level of agreement between observers and teachers reported by these two analyses of reports on subject matter content is substantial but not spectacular. Burstein, McDonnell, and their co-authors identified some patterns within the varying levels of agreement on the mathematical topics covered within their study. Topics in upper-level high school courses were reported with greater accuracy, as were reports

on more specific mathematics topics. Subject matter used as tools in the teaching and learning of other topics – e.g., using tables and charts to record measurements of geometric figures – was reported less accurately. So, significantly, was content associated with recent mathematics reform. The authors write: "the lack of common agreement on the meaning of key terms associated with the mathematics reform movement (e.g., math modeling, patterns and functions) is likely to result in misinterpretation of the data." (Burstein et al, 1995, p. xii). These trends are intriguing, and warrant further research.

Method

This paper relies upon an interpretive approach to research and theory building (Geertz, 1973; Glaser & Strauss 1967). It investigates differences in log reports of classroom practice by exploring the meanings different actors – log developers, trained observers, teachers – assign to particular words. This work borrows theories and techniques from social theorists and linguists, who argue that meanings are not fixed and immutable but variable within and across communities that use particular words (Bakhtin 1981; Freeman 1993, 1996; Gee 1999). Gee (1999) argues for using such situated meanings as tools of inquiry, and we do so here to illuminate ways in which survey instruments become problematic when used across communities.

Data collection for this study took place in eight elementary schools during the spring of 2000. These schools were each implementing one of the whole-school reform programs being tracked by the [name of project], which is investigating the design and

enactment of three leading whole school reforms[1] and their effects on students' academic and social performance. Because [name of project]'s ambitious data collection effort (120 elementary schools and roughly 20,000 students, over six years) necessitated valid and reliable indicators of school and classroom processes, [name of project] engaged these eight schools to help pilot and improve survey instruments.

One such survey instrument is a log of mathematics instruction (see Fig. 1). Teachers participating in [name of project] complete the log daily for six-week periods three times a year, reporting on both the subject matter content and instructional approach delivered to a randomly generated "target student" in their mathematics class. To help determine how to complete the log, teachers in both the spring 2000 validation and our main study attended a day of training on the use of this and a related English Language Arts instructional log; they can also refer to a glossary, which describes the subject matter content or instructional practices entailed by each item; call a toll-free hotline, where questions can be asked of [name of project] staff; and refer to a site facilitator, who can answer non-content questions (e.g., "should I also log my 'calendar mathematics' period?"). In most cases, the glossary gave definitions and examples for each item. However, these examples were far from mathematically complete definitions, and were also far from an exact description of how particular classroom activities mapped to log terms.

[1] The interventions were Accelerated Schools, America's Choice, and Success for All. Community for Learning schools participated in the pilot, although this program is not now part of the main study.

As part of the validity study, pairs of observers watched 29 teacher-participants deliver mathematics instruction for one day's lesson. After the completion of the lesson, the two observers in each classroom wrote detailed fieldnotes with as many verbatim quotes as possible, and both the observers and teacher logged the instruction delivered to one target student. Focusing on one target student prevented the confusion that might arise when teachers individualized instruction, had students working in small groups, or otherwise differentiated instruction. It is these 29 sets of logs that form the basis for the match rates presented in the results section. Observers reconciled differences in their log among themselves, then one observer interviewed the teacher about the log record of that day's lesson. Differences among observers were recorded and reflected upon in writing; teacher-observer differences were explored in the interviews. In most cases, the challenge to observer and teachers was to find out why disagreements arose: what part of the lesson led someone to mark an item? What interpretation of the item made it representative of a teachers' practice? Text generated through interviews and reflection forms the basis for qualitative analysis presented below.

Because of its content specificity, with a significant focus on the teaching of particular mathematics topics, the log provides a fruitful site for learning more about the validity of terms meant to represent mathematical content (see Fig. 1). A "gateway" series of items, for instance, asked teachers whether certain mathematics topics had been taught and if so, with what emphasis. In order to obtain additional detail about some

lessons, the log asked teachers who reported that they taught certain "focal topics" -- counting or ordering, place value, fractions, or multi-digit operations -- to answer more detailed questions about that particular topic. For these focal topics, log users reported only the presence or absence of a particular topic or activity.

Decisions about the design of the gateway and choice of focal topics for this piloted log were made in response to multiple concerns. Developers were aware of national standards and content breakdowns, including those used in the early drafts of the Principles and Standards for School Mathematics (NCTM 2000), and in older versions of both NCTM and other frameworks. Yet this log was the fourth one piloted by [name of project], and feedback from teachers about previous versions often disrupted the neat categorization of strands in national documents. Problem solving and communication and representation were added to the gateway section after many teachers reported that these topics were not mere processes, but the intended focus of mathematics lessons. Since length was a concern, the number of "focal topics" was limited to four. Finally, log developers also wanted data to answer research questions, criteria which helped guide the choice of focal topics toward central topics in K-6 mathematics instruction.

Teachers and observers could mark multiple gateway and focal topic items to indicate lesson content. A lesson on adding fractions which touched upon algebra and engaged students in developing and evaluating conjectures, for instance, would be logged at

operations with fractions, algebraic reasoning, and exploration and problem solving. In practice, however, observers tended toward parsimony, choosing to represent lessons with as few checks as possible. Fractions and multi-digit operations comprised the bulk of the detailed information available from this study, as very few lessons observed featured place value or counting and ordering, the other two focal topics, prominently. For this reason, this analysis focuses on these two areas in depth, as well as data from "gateway" items.

All but four teachers participating in this study were white females. The distribution of teachers across grades is shown in Table 1. Because observers, and the perspectives they brought to the validation study work are also important to this analysis, we present some information about them. Of seven observers, five were doctoral students in an education studies program. Of those five, one had been an elementary teacher, another three had tutored or taught in special education settings, and another had taught graduate students in education. The other two observers had doctorates in other social science disciplines. Observers attended 24 hours of training on the use of the log. The log was written by [colleague], [author], and others working to refine content and instructional distinctions, including experts in both survey research and mathematics. Log authors thus worked across the communities described below – some had recently been mathematics teachers, others were involved in mathematics education improvement efforts and reform, others were survey researchers, and several log reviewers were research mathematicians.

Because observers did not hold "the" authoritative view of correct log usage, and because these data allow a look inside classrooms by reading and analyzing two observers' narratives of each mathematics classroom, this paper contrasts three perspectives: that of observers, teachers, and log developers. The method was simple: ignoring items that focused exclusively on teaching practices (discussion, explanations) and the cognitive demand of students' tasks, we searched for explanations for divergences in observers' ([author], among others) and teachers' application of subject matter terms to particular instances of classroom instruction. This search was aided by the fact that observations, interviews, and observer's reflections had been entered into NUD*IST, software designed to allow the management and analysis of qualitative data. For any given disagreement, an analyst (the author, and other log developers) could triangulate relevant written observations, reflective comments from observers, and teachers' interview transcripts. Log developers did not independently code observers' notes, focusing attention instead solely on cases of disagreement between observers and teachers, or observers and observers. The original analysis which led to this paper took place in early summer of 2000, as [name of project] was adapting the log to its final form.

Results: Validating Mathematical Content Measures

Match rates for the mathematics log were calculated in three ways for gateway items: first by calculating the overall rate of exact agreement, including cases in which the

teacher and two observers each indicated that a mathematical topic did not appear in the lesson, or what we call "zero-zero matches"; second by calculating the exact agreement rate excluding the zero-zero matches; third by calculating a match rate which excludes zero-zero matches, but includes "off-by-one" matches, similar to RAND researchers' method. This last method calls any two adjacent categories – a 1 and 2, or 3 and 4 – a match. As Table 2 shows, the "exact match excluding zero-zero" category is the most stringent criteria; "exact match including zero-zero" is the least stringent[2]. Table 3 shows similar calculations for focal topics, where rates were only calculated given a member of the teacher/observer trio entered the section, and where off-by-one rates were not calculated due to the binary nature of logged reports. The rates either show impressive agreement or room for improvement, depending upon how one views zero-zero-zero and off-by-one matches.

Insert Tables 2 & 3 here

Explaining disagreements

To investigate the causes for disagreements, this investigation coded text describing such disagreements in NUD*IST. Single disagreements were occasionally coded more than once, as when an observer disagreed with the teacher for one reason, and the other observer for a different reason. Roughly half of all disagreements could not be coded, either because observers' comments or teachers' interviews did not contain

[2] Including zero-zero matches also tends to privilege items which received relatively less overall use, such as number patterns, functions, and inequalities, since the more "zero" matches, the higher the overall match rate

enough information to make a judgment as to the cause of the divergence, or because the disagreement appeared to exemplify its own category, leaving the analyst without a method for determining whether it was part of a pattern or simply a random event.

The coding system developed through an iterative process. Based on an initial exploration of the data, analysis began with a set of categories which reflected the location of the disagreement – in individual cognition, social structures, etc. Results of this coding are presented below as these themes in both quantitative (proportions) and descriptive form; the proportions should not be thought of as frequencies generalizable to other instruments of this kind, but instead an extremely rough indicator of the relative frequency of problems identified here. These problems fall into four categories: memory/perception, how mathematics is taught in elementary classrooms, meanings and language, and instrument design. We will focus most intense scrutiny on the third, since it is a novel explanation for measurement error.

Memory/perception.

Although daily logs are used as one remedy for cognitive failures, problems of recall and perception explained about 15% of the overall disagreements coded. In such instances, an observer or teacher indicated they "should have" marked a log item, or that they "forgot" about an event or topic covered in class. Observers and teachers were most likely to pass over events which were brief moments in the context of a whole day's mathematics instruction – for instance when a student completed one fraction problem on a worksheet filled with many different types of problems. Coders'

reflections also occasionally indicated disagreement over whether such brief events even occurred, suggesting that one or more parties failed to perceive the event at all.

How mathematics is taught in elementary classrooms.

The analysis of pilot data revealed problems that arise because of the nature of this subject matter and how mathematics is taught. For example, in some lessons students touched on one log category in the service of learning another – as when students reduced fractions in the course of performing operations with fractions, or used addition basic facts while solving a multi-digit multiplication problem. The instrument, at the time of the field trial at least, did not explain how to deal with this kind of event.

Another problem which arises from the nature of classroom mathematics instruction stems from the potentially large set of organizational schemes which might be used to represent content. There is no definitive ordering and mapping of the elementary mathematical terrain, which left us to make decisions about how to construct categories teachers would find easy to remember and use. Wanting to include a relatively limited (perhaps 20) set of mathematical content topics on the log, we were thus constrained in this task by needing to subsume finer-grain topics within broader categories. But this led to problems. For instance, many teachers assign students tasks which require representing data through graphs, tables, or charts. Wanting to place this activity under a broader category left us with a choice: place it under "communication and representation," or place it under "statistics." Arguments could be made for both

categorizations, yet we eventually chose one – statistics – which caused at least one disagreement on the log. Similar problems occurred where the mathematical point of an activity was not clearly stated or potentially multiple; skip counting, for instance, can contribute to students' ability to count, but also their ability to remember basic facts, identify patterns in number, or even begin to learn about functions. Our glossary, however, placed skip counting under counting.

Another set of problems in this category arose because of the way mathematics is currently conventionally taught in U.S. classrooms. In contrast to older constructions of mathematics lessons, which tended to focus on one topic (e.g., 2-digit multiplication; inequalities) at a time, most contemporary curriculum materials contain not only the new material for the day, but a smattering of other topics intended for students' review and practice. As one observer recorded, students worked in one class on a mathematics worksheet which included "the following assortment of problems: an average; a finding the difference in years word problem; a percent to fraction; multiples; writing a shaded part of a grid as percentage, decimal, and fraction; shape of a basketball; changing 1 ½ years into months; area and perimeter; multiplication of decimals by 10 and 1000; adding, subtracting, and multiplying decimals; adding fractions and mixed numbers, subtracting mixed numbers; multiplying and dividing fractions." Although this lesson is on the extreme end of the spectrum, such spread of topics was common in the classrooms observed. It also complicated logging, increasing the possibility of a log user forgetting a topic covered, and raising in many observers'

minds questions about how much emphasis on a topic was required before marking it as a focus of daily activity. This genre of mathematics lesson also raises problems for analysts, as well; if the activities in the lesson like the above are each recorded, it likely over-emphasizes students' exposure to such topics, relative to more conventional lessons where such topics are accorded at least a few minutes' time. If the activities are not recorded, the log misses student practice on these problems.

Finally, students' independent work and practice, another feature of some mathematics classrooms, decreased levels of agreement in teachers' and observers' records. In at least one school participating in this study, students practiced mathematics problems on computers during a separate instructional period. In another school, students completed "wait time" activities – time spent between teacher-led activities doing dittos and worksheets. In neither case did the teachers in these classrooms have close knowledge of the content target students covered during these periods. Observers, who had the luxury of focusing on one student only, did have such knowledge, often leading to mismatches in topics logged.

Although we cannot claim that the 29 teachers observed for this field test engaged in completely typical U.S. elementary mathematics instruction, they did represent a range of degree of engagement with improvement efforts, with some using novel curriculum materials associated with whole-school reforms, and others using quite traditional materials (e.g., Saxon). Thus we argue these mismatches suggest those which might

arise in a more representative sample. In all, roughly a third of disagreements in the log pilot arose because of the interaction between the way mathematics is taught in today's classrooms and our efforts to capture that work.

Meaning and language*.*

There is no absolute, fixed relationship between any word and its meaning(s). What we call a "chair" might have easily been called a "table" instead, and vice versa. Instead, meaning is assigned to words by the communities who use them. Different communities may assign different meanings to particular words or phrases – for instance, "barbeque" means one thing in South Carolina, and something else in Georgia. By extension, different communities may also impute different meanings to the same phrase, and may also differ linguistically in other ways – the precision with which particular words refer to objects or ideas, for instance, or the grammatical structures used. We refer to this idea as the way language is used, or "language use," in particular communities.

There is evidence that professional communities constitute important units of analysis in language use. That is, social theorists and education scholars have written about the specific ways meaning-word relationships are constituted within and travel across professional boundaries. Bakhtin (1981), for instance, distinguishes between "professional" and "generic" languages (p. 272; 293). Freeman (1993; 1996), Lampert (1999) and others study individuals' use of language to understand teachers' journeys

from novice to professional, insider to outsider, apprentice to expert, or from local communities of meaning to more remote. And Jackson (1968) has written about the lack of "technical vocabulary" in teaching.

In this case, we argue that three different professional communities give meanings to the terms on the log. These communities are comprised of a) non-teaching observers, some of whom were survey researchers b) elementary teachers and other practitioners, and c) mathematics educators and researchers engaged in efforts to improve elementary teaching. An examination of mismatches reveals that almost a third of disagreements were explained differences in meaning across these community boundaries. Below, we describe the nature of language use in these fields, and patterns we found in the data. Although we draw here on both basic ideas from linguistics and some available evidence regarding the history and nature of language use in these three fields, the reader should understand these arguments as arguments, constructed for the sake of provoking discussion and thought. Grouping particular terms with particular professional communities, for instance, must be thought of as informed conjecture, rather than a matter of scholarly record.

Access to mathematical language

Evidence suggests mathematicians have a specialized language for communicating about subject matter content. To start, mathematicians are famously particular about how subject matter terms are used, eschewing ambiguity for precision and the tightly denotative use of terms. In some degree, the care with language helps explain

differences between the ways individuals use what some observers call "natural" or "ordinary" language and the ways in which mathematicians use language (Smith 2002; Pimm 1987). Where "natural" language often contains terms with multiple meanings, mathematical language is much less likely to do so; in addition, the same words often have separate mathematical and natural meanings. Some also report that the process of becoming a professional mathematician involves an apprenticeship in which initiates come to understand terms, grammar, and modes of discourse through using them in the practice of mathematics.

Relatively high levels of agreement obtained for log terms historically associated with university mathematics, perhaps because of the nature of mathematical language within the university. Log users tended to agree on log items which used such terms, e.g., geometry, probability, statistics, and functions. An examination of the disagreements in this category, further, showed a pattern: disagreement occurred when log users' interpretations of mathematical terms varied from the interpretations given by the discipline of mathematics. For instance, one observer logged a lesson on representing student preferences for ice cream flavors by marking "percent, ratio." While this lesson included an emphasis on representing data, students did not explicitly discuss ratio or percentages, or represent ice cream preferences as percentages or ratios. Observers and teachers both had difficulty with the item "justification and proof," categorizing these instances of students sharing how they found their answers (e.g., "we counted up by ones") and other events that mathematicians would perhaps

not classify as proof ("I checked with a calculator"; a show of hands in the classroom to determine whether a particular mathematical point was true).

In these examples, the data suggest that log users were not aware of or able to use mathematical definitions and knowledge to classify classroom practice. From one perspective, we might simply argue that accuracy in survey research in mathematics is related to survey user's mathematical knowledge; certainly the percent/ratio example suggests this. However, the justification and proof example also suggest viewing this through the lens of users' access to mathematical language. No observer or teacher provided evidence of strong ties to the discipline of mathematics, suggesting this particular set of log users had limited access to the mathematical meanings intended for this item. Without such access, log users substituted (or constructed) everyday or "natural language" definitions for terms.

Access to elementary teachers' language

We argue that the community composed of elementary teachers and others (in particular, authors of some curriculum materials) has imbued terms like "inequalities" "basic facts" or "equivalent fractions" with particular associations, associations that refer to mathematical content. "Basic facts," for instance, typically refers to computation in which at least one number in the posed problem has only a single digit – e.g., 17 – 9 = 8, but not 17 – 12 = 5. These associations may not only entail subject matter content but also knowledge of particular student and instructional tasks. Sherin, for instance,

found knowledge of particular topics was linked to instructional tasks and ways of teaching the topics (Sherin 1999). The same might be true of elementary school teachers. "Equivalent fractions" typically refers to efforts to show or find fractions that represent the same quantity – 1/2 and 2/4, for instance. Although fractions can be equivalent in other contexts, for instance when reducing or comparing fractions, teachers might limit their use of this descriptor to cases in which students are working explicitly on the idea of equivalence. Textbooks, standards, and professional developers often organize and name lessons with these terms, and it is likely that teachers could communicate with some precision about student and content by using such terms. Mathematicians might not use such terms in these ways, but they constitute accepted ways of talking about mathematical content in elementary schools. This "school mathematical language" may even facilitate communication and cooperative work amongst teachers, in the way that technical language in law or medicine facilitates professional communication in that field.

Evidence of this language around practice comes from instances in which observers, particularly those without prior experience in elementary mathematics classrooms, interpreted log terms in non-standard ways. "Ordering fractions," for instance, typically refers in curriculum and teachers' language to activities which ask students to place three or more fractions in ascending order – e.g., 'put 1/2, 1/8, and 3/4 in order." One observer, however, used this term to refer to counting with fractions (e.g., 1/2, 1, 1 1/2, 2). Though the result is, literally, ordered fractions, it is not what is meant by

teachers and others who use the term "ordering fractions." Likewise, in several instances observers indicated they were unsure whether particular problems (e.g., "20 + 4" "12 – 5 = 7") fell into the "multi-digit" or "basic facts" category; teachers, perhaps more used to the conventions around such definitions, did not question this term. Finally, the term "inequalities" refers to tasks which ask students to determine whether one number is larger than another:

104 ◯ 14 1/2 ◯ 2/5

One observer, however, used "inequalities" to code an activity in which students identified food broken into equal and unequal parts. Although technically dealing with inequalities – the unequal parts – this term is, we argue, conventionally used to refer only to the above mathematical activity and instructional content.

Observers, most of whom had not been classroom teachers in elementary schools, cannot be faulted for failing to recognize the conventions associated with such terms. They had little or no access to the language used by teachers and others to describe elementary mathematics classrooms, and how the teaching community marked the occurrence of particular events or activities. However, the effects of their lack of access to this use of language proved illuminating, for it helps provide emergent evidence for a language of practice within elementary school mathematics. Whereas scholars have long been skeptical of the existence of a professional language within education (Jackson, 1968), teachers' use of these categories generally accorded with what log writers intended, even though observers' did not. This indicates some shared

interpretations of particular terms, and what practices instantiate those particular terms. Only when outsiders entered classrooms, and used lay definitions to interpret these mathematical terms, did this language become visible.

Access to language used by mathematics reformers

In recent years, a community of mathematics educators comprised of researchers, teaching faculty, mathematicians, policy-makers, curriculum developers, and others has emerged around efforts to improve mathematics teaching in U.S. classrooms (e.g., NCTM 1989, 2000; NRC 2001). We argue that like the mathematics and elementary teaching communities, this reform community has its own language to facilitate communication about subject matter content. We argue that many of the terms used by this community have variable meanings – sometimes within the community itself, as it comes to agreement on the meaning of particular terms, but more often as terms from this community are used by members of others.

To communicate its vision, the mathematics reform community has both redefined older terms and coined new ones. An example of redefinition can be found in "problem solving," a term with a long history of meanings, but which is now used by many in this community to represent the work students do when they puzzle over an unfamiliar and difficult problem (see Schoenfeld 1985; 1989). When used in this way, as it was intended to be on the log, it represents the process students engage in as they grapple with a problem for which they "(do) not have a readily accessible mathematical means

by which to achieve resolutions," (Schoenfeld 1989, p. 88). Many teachers reported problem solving of this sort might also be taught explicitly, as a mathematical topic in its own right. In the wider domain of schools, curriculum materials, and professional development, however, the term "problem solving" is often used to represent more conventional practices, such as solving word problems for which the solution method is fairly obvious, or introducing students to new manipulative materials. Thus it is not surprising that, despite a glossary definition for this term, teachers and observers used it to represent both these more conventional activities as well as more difficult student tasks more in line with our original intent.

When mathematics reformers introduced new terms, difficulties also arose. For instance, frequent disagreements arose around the following categories:

- Steps of a standard procedure or algorithm[3]
- Transitional forms of the standard procedure—e.g., using partial products in multiplication
- Alternative or non-standard methods for solving multi-digit computation

Most observers of U.S. classrooms report that the first activity, the steps of a standard procedure or algorithm, is the most common method for teaching students multi-digit computation (see Fig. 2 for the compact algorithm for multiplication). Yet newer curriculum materials (e.g., Everyday Mathematics, Investigations) employ transitional

computation procedures to make plain the mechanics or meaning behind a multi-digit operation. Fig. 2, for instance, shows the transitional procedure for multiplication, which involves finding each partial product (e.g. 6 x 8, 6 x 20….), then adding them up. This method better identifies place value in each step of the algorithm, and allows for potential comparisons between the compact and this expanded method. By making such a comparison, students might understand the compact method more thoroughly. Finally, alternative algorithms (also referred to as "non-standard") are not developmental, but simply other algorithms for solving computational problems. In Fig. 2, the third version of 56 x 28 shows an algorithm which inverts the usual procedure, beginning by multiplying 8 x 6, carrying the four, then multiplying 8 x 50, etc.

Insert fig. 2 here

Our data suggested that despite explanations in both the mathematics glossary and during observer training, log users' lack of knowledge about these reform mathematics terms complicated their reporting. For instance, one teacher explained why she chose the "transitional form" item to represent her lesson: "(after a pause) okay, how we broke the process down and how we talked about the steps and how important it is if you skip a step or you even skip one of the multiplication, if you don't do your facts, the whole process is not going to be right. We set it up in stages and we do it in order." Log observers' records show that this teacher emphasized the steps in the standard, compact U.S. algorithms during her lesson. However, the text above suggests she took that careful emphasis on the steps (… "we set it up in stages and we do it in order…")

[3] According to Bass (2003), an algorithm "consists of a precisely specified sequence of steps that will lead

to mean that she was using a transitional form of that procedure. She ignored the item's cues (and unfamiliar terms) about "partial product" and "transitional forms." These terms proved problematic for observers, as well; one wrote "the method they used to solve multi-digit computations was rather unorthodox, but my mathematical knowledge is too weak to evaluate whether this method is (an) alternative or non-standard (procedure)." Thus terms used by those seeking to improve mathematics education are not yet familiar to those outside this community.

Other log terms suffered, in terms of accuracy of use, from similar problems. One teacher reported that she had difficulty distinguishing whether an activity involved number patterns or functions. In another case, a task which asked students to identify a pattern in a geometric representation (towers which added an increasing number of blocks for every subsequent row) was marked by one observer as a geometric pattern, another observer as a numeric pattern, and the teacher as exploration and problem solving. Cases can be made for all these categorizations of this activity, yet the three individuals represented it using three separate items. Observers also puzzled over what to label as use of "concrete models" (e.g., do slash marks written on a blackboard count?). In many such cases, the glossary provided with the instrument did not help, ironically adding only more ill-defined terms to the process.

In all, terms newly defined or redefined by mathematics reformers (number patterns; geometric patterns; communication and representation; exploration and problem

---

to a complete solution for a certain class of computational problems."

solving; meaning of numerator and denominator; meaning of part-whole ratio with sets; representing fractions or equivalence with concrete materials; connecting two or more concrete representations of fractions or equivalence; connecting concrete representations of fractions to number and symbols; why procedures work)[4] have only a 40% match rate (excluding zero matches), as compared to 68% for terms which we argue have relatively more stable and widely known definitions within mathematics or conventional elementary mathematics communities. Table 4 shows examples of terms from each category.

We argue the depressed match rate for this set of terms derives from log users' lack of access to the language used by the mathematics reform community described above, and perhaps from the imprecision with which some of these subject matter terms have been defined in the research literature. The first explanation points to the difficulties entailed in having "conversations" (Gee, 1999) across community boundaries. Years of experience have led mathematics reformers to shared understandings of particular terms. Teachers and observers, most of whom had not participated in that learning process, had no means to reach common understanding of those terms. The second explanation suggests that even within the mathematics reform community, some terms remain undefined. Some who have read this article, for instance, disagree with the field tested log's characterization of transitional and alternative algorithms. In this case,

[4] Several items, including algebraic reasoning, transitional and alternative methods for solving multi-digit computation, etc., received fewer than five uses and were not included in the quantitative analysis. We also recognize that ambiguities exist in this list, and disagreements may occur over where to place

the possibility for accurate measurement of subject matter content is further diminished.

Instrument design.

Finally, teachers' and observers' efforts to use the log to record classroom practice were complicated by the design of the instruments, notably the glossary. The glossary was relatively short; it relied more on bulleted examples and references to other mathematical terms than it did to longer, more in-depth explanations of the mathematical content. Providing more detail might have alleviated some of the problems associated with language and meaning, although it could not possibly have eliminated them. While a glossary provides users some context for terms, it cannot provide access to the ways particular communities imbue terms with meaning. Further, glossaries are static collections of words describing other words, rather than concrete instantiations or uses of such words in the practice of doing mathematics. Finally, the glossary also contained confusing definitions for a number of terms. For instance, "ordering fractions" included, inexplicably, reducing fractions, equivalent fractions, and like and unlike denominators in its definition. Roughly one-tenth of disagreements arose from problems with the glossary.

Discussion & Conclusion

particular items. However, this list was constructed in consultation with both mathematics educators and mathematicians, and represents the best attempt at such a classification system.

We cannot be sure how the patterns identified here generalize to other grade levels, subject matters, or even other studies of the use of mathematical terms in classrooms. Certainly, more research is needed into the use of mathematical terms in elementary schools, into the processes by which teachers and observers connect concrete events and activities with broader mathematical categories, and into why errors in reporting arise.

However, the results uncovered here suggest some tentative conclusions for those writing, using, and interpreting results from instruments like our mathematics log. Terms used to represent mathematics subject matter content are not shared as commonly as many seem to assume. By examining disagreements closely, this analysis identified patterns within this finding: that log users' knowledge of mathematical definitions and terms affected the accuracy of their reports; that log users' knowledge of the conventions which connect terms and elementary classroom practices likewise affected accuracy; that some terms, particularly those more recently inserted into the conversation about mathematics reform, proved to have particularly low rates of shared interpretation and agreement. We argue for viewing these difficulties not simply as problems with log users' knowledge or terms' definitions, but also as problems inherent in designing survey items with common meanings across different communities. This analysis also identified the ways in which structural features of elementary mathematics education, as taught currently, affect agreement and the logging process itself. And

this analysis confirmed the presence of conventionally understood sources of error, such as memory, perception, and instrument design.

This analysis carries several practical implications. First, these problems remind researchers that measurement error arises not only from conventional sources, but also from characteristics of the terrain meant to be measured. Measuring mathematics content, for instance, may be more difficult than measuring teachers' use of different grouping arrangements or the number of students who attend class on a given day. This adds an additional – and difficult to predict – complication to instrument designers' work.

Second, the confirmation of RAND researchers' findings (Burstein, McDonnell et al, 1995) that "reform" content (e.g., number patterns) is less accurately reported than "traditional" content raises some concern. Researchers have come to expect such error, and have sophisticated methods for coping with it, but when some of that error is unevenly distributed over content items, it can pose serious problems for independent variables in many statistical techniques. The more error and the less "true" signal compose an independent measure, the more it resembles a stochastic (or random) variable, and the less likely it will appear systematically related to any dependent measure – even if the "true" relationship is strong. Said another way, measurement

error, at least in bivariate regressions, tends to reduce the absolute value of the estimated coefficient (Hanushek & Jackson 1977, p. 288).[5]

This poses a problem for analysts interested in using error-prone instruments to predict school or student performance. Analysts have, for the most part, dealt with this problem by emphasizing significance levels, rather than the absolute size of a coefficient, when describing a variable's effects. The problem becomes more complex, however, when analysts wish to know something about the comparative value of two theoretically separate independent influences upon an outcome. Intuitively, it seems important that the hypothesized independent influences have roughly the same size error component proportionate to the size of the "true" signal. This can be demonstrated through the problem which arises if the actual effects of two variables were roughly equal, but one were composed of mostly "true" signal, and another mostly "error." In this situation, the more accurately measured independent variable would appear more effective than the latter, despite the similarity in their actual size. This situation might apply, for instance, to linking student outcomes with engagement with either conventional mathematics topics – improper fractions, proper fractions, operations with fractions – or newer topics such as problem solving, justification and proof, or mathematical communication. More disagreement about what those novel terms means such topics are likely to be less accurately measured, and may be less likely to show an effect in statistical models.

[5] The situation for multivariate regression is more complex; see Hanushek & Jackson 1977, p. 288.

Our findings also offer a theoretical lesson about language, and the role it plays in both schools and research efforts within and around schools. Over 30 years ago, Jackson (1968) observed that "one of the most notable features of teacher talk is the absence of a technical vocabulary" (p. 143). More recently, others have studied language use around recent education reforms, reporting again that a lack of common language among and between policy-makers, teachers, teacher educators, and scholars hampers efforts toward those reforms (Author, 2001; Author & Colleague 2000). The findings here slightly modify and expand, to some degree, upon these studies. Teachers and log developers did appear to have some common language for communicating about the content of mathematical work in classrooms – a school mathematical language; the existence of this common language became clear as individuals inexperienced in classroom mathematics attempted to use the instrument. The extent of this common language, however, has yet to be mapped out. While well-established terms like "inequalities" and "ordering fractions" were commonly used with shared meaning, many others were not.

Viewing language use across the communities provides another perspective, one which footnotes this last conclusion. In contrast to the unwritten system of conventions which links particular terms to mathematical content and practices within U.S. classrooms, mathematicians have a longer tradition of using formal definitions as part of their everyday work. Unlike natural language, which allows considerable variability in meaning, and in which connotation is often important, mathematical language is

carefully and precisely denotative. Definitions matter. Mathematical definitions are explicit, unambiguous, and consistent (Smith, 2002). This care with terms allows mathematicians to communicate clearly and understandably about mathematics, to assume shared meaning, and to refer clearly to particular objects, actions, and ideas. This analysis suggests that some of the language around elementary mathematics education lacks both the precision with which mathematicians speak about mathematical topics and content, and carefully crafted definitions and meanings for particular subject matter content. This is a problem for research, but it might also pose problems among those who work in classrooms, with students, and who daily interact over how a particular student is doing or whether particular content has been covered.

This point is also illustrated by an episode which occurred as analysts sought to re-write the log. As the log was revised, its authors noticed that it contained terms and phrases that stem from the prevailing school curriculum, but which are not part of conventional disciplinary mathematical usage. For instance, one portion of the log asked teachers whether they covered "mixed numbers;" another asked teachers whether they covered "decimal numbers." These terms do not refer, except indirectly, to classes of numbers, but rather to notational representations of them. For example, the same number, one and one quarter, can be written (or represented) as a decimal (1.25), as a mixed number (1 1/4) or as a fraction (5/4). Thus, 1 1/4 is a "mixed number," while 1.25 and 5/4 are not, though they have the same numerical value as 1 1/4. Mathematicians tend to pay less heed to naming notational representations of numbers, and focus instead on

classes of numbers themselves, for example the natural numbers (1, 2, 3, …), the whole numbers (0, 1, 2, 3, …; add zero to the natural numbers), the integers (…, -2, -1, 0, 1, 2, …; the whole numbers and their negatives, or additive inverses), or the rational numbers (numbers expressible as the result of dividing one integer by another one different from zero, or p/q with p, q integers and q not equal to 0). Thus 1.25, 1 1/4, and 5/4 all represent the same rational number.

Thus the language of practice used in elementary mathematics identified earlier is not a mathematical language, in the sense that it shares meanings and conventions with the language mathematicians have constructed. This observation is shared by others. Jim Lewis, a mathematician who is currently engaged in teaching mathematics to preservice teachers, writes:

> In my world one talks about "natural numbers," "integers" and "rationals." In the world of the elementary school teacher one talks about "whole numbers," "decimals," "fractions," "negative numbers" etc. I've discovered this year that the distinction makes it very hard to communicate with the students in my class. (Lewis, personal communication 3/30/01)

Evidence from recent efforts to improve mathematical teaching and learning also suggests a disjunct between everyday and mathematical languages for talking about classroom mathematics instruction. Judith Roitman (1998), a mathematician considering the 1989 NCTM standards, wrote:

> Although I am generally pleased by the major directions of the standards, it is undeniable that the standards documents are peppered with statements that are mathematically questionable. Generally, these are not anything as simple as a straightforward mathematical mistake. Their best description is as something no one who really knew the mathematics would say. (p. 28)

That teachers and mathematicians speak different languages is not surprising, for different professional affiliations bring different training, norms, customs, and knowledge. However, there are reasons to unite the language used by these communities. One is some scholars' call for children to be doing mathematics as mathematicians do it (see, e.g., Lampert 1990); if such a move toward authentic work within a discipline is to occur, the language used to support such work needs to reflect that used by mathematicians themselves. Another is credibility within and across communities with stakes in mathematics education. If mathematics educators or researchers are sloppy in the use of mathematical terms, it casts reasonable doubt on the carefulness of our inquiry and analyses. Recent critiques of reform movements (e.g., ones made by the group Mathematically Correct) have focused on such inaccuracies. Finally, the use of language is educative; if instruments like our mathematics log, or classroom discourse itself support inaccurate usage, students will have more difficulty learning content.

Table 1: Teachers' grade level

Grade	Number of teachers
First	6
Second	5
Third	11
Fourth	4
Fifth	3

Table 2: Agreement rates for gateway section of log

Gateway item	Exact matches, including zero-zero	Exact matches, excluding zero-zero	Off-by-one matches, excluding zero-zero
Basic fractions	0.83	0.08	0.42
Decimal fractions	0.94	0.50	0.50
Ordering fractions	0.90	0.45	0.82
Improper fractions or mixed numbers	0.90	0.25	0.75
Operations with fractions	0.89	0.22	0.56
Multi-digit addition and subtraction	0.73	0.43	0.90
Multi-digit multiplication and division	0.83	0.31	0.63
Addition and subtraction basic facts	0.67	0.34	0.75
Multiplication and division basic facts	0.87	0.73	0.87
Number patterns	0.84	0.23	0.62
Percent, ratios	*	*	*
Geometry	0.78	0.22	0.72
Geometric patterns	0.90	0.00	0.50
Measurement	0.76	0.40	0.84
Probability	*	*	*
Statistics	0.87	0.11	0.78
Functions	0.92	0.44	0.56
Inequalities	0.89	0.00	0.00
Algebraic reasoning	*	*	*
Mathematical communication and representation	0.40	.07	0.51

Exploration and problem solving	0.54	0.15	0.50
Justification and proof	0.63	0.08	0.36

* Indicates less than five uses by teachers or observers.

Table 3: Agreement rates for focal topics sections of log

Fractions	Match rate including zero-zero	Match rate excluding zero-zero matches
Meaning of numerator and denominator	0.857	0.500
Meaning of part-whole ratio with sets	0.857	0.333
Meaning of part-whole ratio with regions	*	*
Meaning of fractions as division of two whole numbers	*	*
Meaning of fractions as points between whole numbers on the number-line	*	*
Equivalent fractions	0.857	0.667
Comparing size of fractions	0.714	0.500
Ordering fractions	0.714	0.500
Representing fractions or equivalence with concrete materials	0.571	0.500
Connecting two or more concrete representations of fractions or equivalence	0.714	0.000
Connecting concrete representation of fractions or equivalence to numbers and symbols	0.571	0.500
Finding common denominators	0.857	0.500
Operations with fractions (adding, subtracting, multiplying, dividing) with concrete materials or pictures	*	*
Operations with fractions (adding, subtracting, multiplying, dividing) in symbolic form	1.00	1.00

Operations

Steps of standard procedures or algorithms	0.875	0.857
Why a standard procedure or algorithm works	0.500	0.000
Regrouping-e.g., ones, tens hundreds; and tenths, hundredths, etc.	0.750	0.714
Transitional forms of the standard procedure—e.g., using partial products in multiplication	0.500	0.000
Alternative or non-standard methods for solving multi-digit computations	*	*
Why an alternative or non-standard procedure works	*	*
Connecting a concrete model to the steps of a procedure	*	*
Comparing different methods for solving multi-digit computations	*	*
Using computational procedures to solve problems	.625	.625

* Indicates three or fewer uses by teachers or observers.

Table 4: Examples of terms used in [name of project] log pilot

From mathematics community	From elementary mathematics teaching community	From mathematics education and research community
Geometry	Inequalities	Algebraic reasoning
Fractions	Ordering fractions	Number patterns
Functions	Multi-digit addition	Exploration and problem-solving
Proof	Measurement	

Figure 1: Log "gateway" and fractions section

Figure 2: Compact, transitional, and alternative algorithms

Transitional	Compact/standard	Alternative
56 × 28 48 400 120 1000 1568	56 x 28 448 1120 1568	56 x 28 168 1400 1568

References

Author (2001)

Author & Colleague (2000)

Author & Colleague (2001)

Bakhtin, M. M. (1981) The dialogic imagination: four essays. Austin : University of Texas Press.

Bass, H. (2003). Computational fluency, algorithms, and mathematical proficiency: One mathematical perspective Teaching Children Mathematics, 9 (6), 322-327.

Burstein, L., L.M. McDonnell, J. Van Winkle, T. H. Ormseth, J. Mirocha, and G. Guiton. 1995. Validating National Curriculum Indicators. Santa Monica, Calif.: RAND.

Colleague & Author (2000)

Colleague & Author (2001)

Everyday Mathematics. (1998) Chicago IL: Everyday Learning Corp.

Freeman, D. (1993) Renaming experience/reconstructing practice: Developing new understandings of teaching. Teacher & Teacher Education 9 (5/6), 485-497.

Freeman, D. (1996) "To take them at their word": Language data in the study of teachers' knowledge. Harvard Educational Review 66 (4), 732-761.

Garet, M. S., Porter, A. C, Desimone, L , Birman, B. F., Yoon, K .S. (2001). What makes professional development effective? Lessons from a national sample of teachers. American Educational Research Journal 38(4), 915-945.

Gee, J. P. (1999) An introduction to discourse analysis : theory and method. New York; Routledge.

Geertz, C. (1973) The interpretation of cultures; selected essays. New York, Basic Books.

Glaser, B. & Strauss, A. (1967) The discovery of grounded theory: Strategies for qualitative research. Chicago: Aldine.

Hanushek, E. A. & Jackson, J. E. (1977) Statistical methods for social scientists. New York: Academic Press.

Investigations in Number, Data and Space. (1995). Palo Alto, CA: Dale Seymour.

Jackson, P. W. (1968) Life in classrooms New York, Holt, Rinehart and Winston.

Lampert, M. (1990) When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. American Educational Research Journal 27 (1), 27-63.

Mayer, D. (1999) "Measuring Instructional Practice: Can Policymakers Trust Survey Data?" Educational Evaluation and Policy Analysis 21 (1): 29-46.

National Council of Teachers of Mathematics (1989) Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

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

National Research Council (2001) Adding it up: Helping children learn mathematics. Washington DC: National Academy Press.

Pimm, D. (1987) Speaking mathematically : communication in mathematics classrooms. New York : Routledge & K. Paul.

Porter, A.C., Floden, R. E., Freeman, D. J., Schmidt, W. H. and Schwille, J. (1986). Content Determinants (With Research Instrumentation Appendices). Institute for Research on Teaching. Research series #179. East Lansing: Michigan State University.

Roitman, J. (1998) A mathematician looks at national standards. Teachers College Record (100), 1, 22-44.

Saxon Mathematics. (1995) Norman, OK: Saxon Publishers.

Schoenfeld, A. H. (1985) Mathematical problem solving. San Diego, CA: Academic Press.

Schoenfeld, A. H. (1989) Teaching mathematical thinking and problem solving. In L. B. Resnick and L. E. Klopfer (Eds.), Toward the thinking curriculum : current cognitive research /Association for Supervision and Curriculum Development Yearbook, (pp. 83-103). Washington DC: Association for Supervision and Curriculum Development.

Sherin, M. G. (1996) The nature and dynamics of teachers' content knowledge. Unpublished doctoral dissertation, University of California Berkeley.

Smith, F. (2002) The glass wall: Why mathematics can seem difficult. New York: Teachers' College Press.

Smithson J.L. & Porter, A.C. (1994) Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the Reform Up Close study. CPRE Research Report Series Report #31. Office of Educational Research and Improvement, Washington DC.

Spillane, J. P. & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. Educational Evaluation & Policy Analysis 21(1), 1-27.

Stecher, B. M., Chun, T., Barron, S., & Ross, K. (2000) The Effects of the Washington State Education Reform on Schools and Classrooms: Initial Findings. Santa Monica: RAND. File # DB-309-EDU, 2000.

Stigler, J.W., Gonzales, P.A. Kawanka,T. Knoll, S. Serrano, A. (1999) The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States. Washington, DC: US Department of Education.

Supovitz, J. A. & Turner, H. M. (2000) The effects of professional development on science teaching practices and classroom culture. Journal of Research in Science Teaching, 37 (9): 963-80.