CRESST REPORT 802

KNOWING AND DOING: WHAT TEACHERS LEARN FROM FORMATIVE ASSESSMENT AND HOW THEY USE INFORMATION

JULY, 2011

Greta Frohbieter

Eric Greenwald

Brian Stecher

Heather Schwartz

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Knowing and Doing: What Teachers Learn from Formative Assessment and How They Use the Information

CRESST Report 802

Greta Frohbieter
CRESST/University of Colorado

Eric Greenwald
CRESST/Stanford University

Brian Stecher and Heather Schwartz
CRESST/RAND Corporation

July, 2011

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

To cite from this report, please use the following as your APA reference: Frohbieter, G., Greenwald, E., Stecher, B. & Schwartz, H. (2011). *Knowing and doing: What teachers learn from formative assessment and how they use the information.* (CRESST Report 802). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

KNOWING AND DOING: WHAT TEACHERS LEARN FROM FORMATIVE ASSESSMENT AND HOW THEY USE THE INFORMATION

Greta Frohbieter
CRESST/University of Colorado

Eric Greenwald
CRESST/Stanford University

Brian Stecher and Heather Schwartz
CRESST/RAND Corporation

Abstract

This study analyzed three different middle school mathematics formative assessment programs, examining how features of each program were associated with the information they provided to teachers and the manner in which teachers used the information. The research team found considerable variation in the information teachers obtained from each program and how they used it. They found that greater familiarity with the specific formative assessment system did seem to be accompanied by more integrated use during the school year. They also found that teachers seemed to find it easier to incorporate the systems that had pre-existing assessments than the system that put the burden for assessment design on their shoulders. The results from this study can aide teachers, administrators and other education stakeholders in deciding which formative assessment systems to adopt, planning for the implementation of formative assessment and providing adequate training for teachers, designing formative assessment systems that better meet teachers' needs, setting realistic expectations for the impact of formative assessment systems on a large scale, and lastly, understanding the impact of formative assessment in a particular context.

Introduction

Background

Across the country school districts are being encouraged to promote formative assessment as a powerful improvement strategy, but they lack guidance about what type of formative assessment to implement and how to implement it. The belief that the expanded use of formative assessment will lead to significant gains in student learning is based in part on a review of 250 studies that revealed significant benefits from the introduction of formative assessment in a variety of educational settings (Black and Wiliam, 1998a). In a subsequent article, the authors reported that the effects of these innovations on student achievement were statistically very large compared with those of most educational reforms;

moreover, the authors found that effect sizes were greatest among low-achieving students and students with special needs (Black & Wiliam, 1998b). It is not surprising that these results have captured the attention of many educators.

In addition, formative assessment is getting a boost from policymakers who seek to make education more effective by drawing on lessons from the business sector. These advocates of data-driven decision making argue that educators are not making effective use of the evidence they have regarding student performance. Certain policymakers point to a variety of models for adapting production processes based on information about performance outcomes (Marsh, Pane, & Hamilton, 2006). Assessments should be used *formatively*, they argue, to monitor student performance and permit teachers to adapt instruction to address identified deficiencies (Hamilton et al., 2009; Massell, 2001). This confluence of recommendations has boosted the number of districts that have adopted new assessment systems, which are designed to improve instruction (Stein & Basset, 2004a; 2004b). There has also been rapid growth in the number of commercial products available for this purpose (Burch, 2006).

Despite the widespread enthusiasm for expanding formative assessment, there is still much uncertainty about this strategy. The assessment options being offered under the general label of formative assessment vary widely. Some focus on training teachers to create better assessment opportunities during their ongoing instruction (Wiliam et al., 2004); others provide stand-alone tests to be administered quarterly (see Foster & Noyce, 2004). District choices are further complicated by the fact that few of these approaches have rigorous empirical evidence of effectiveness (Burch, 2006). Moreover, few formative assessment programs offer guidelines about implementation and use. Districts are left to their own devices to determine which kind of formative assessment will be most effective for students, as well as how to train teachers to use the strategy they adopt.

This study addresses the dearth of information on the use of formative assessment and explores the practices of teachers who are using different kinds of formative assessment systems. Moreover, the study specifically examines the kinds of information teachers gleaned from assessment and how they used that information in their teaching. These real-world examples could be useful to districts when formulating decision strategies, implementation plans, and training needs that are related to formative assessment.

Our study sought to answer the following three research questions:

1. What types of information do middle school math teachers acquire about students from a variety of types of formative assessment?

2. How do middle school math teachers use the information provided by these assessments to improve teaching and learning?
3. Is there a relationship between the features of the formative assessment and either the information provided or how the information is used?

What Is Formative Assessment?

The terminology associated with this topic can be confusing. The term *formative assessment* suggests a type of assessment with particular characteristics; whereas, the phrase ~~"u~~using assessment formatively" (or formative use of assessment), implies a particular kind of action on the basis of assessment. This raises the question of whether formative assessment is a type of assessment, behavior, or a combination of the two. The confusion grows when one considers the plethora of related terms that are likely to be encountered in discussion of these issues—including interim assessment, benchmark assessment, periodic assessment, assessment of learning, assessment for learning, data-driven decision-making, data-driven instruction, among others. While it is not the purpose of this paper to explore this lexicon, at a minimum, we should clarify the term *formative assessment* as it will be used herein.

Key Characteristics of Formative Assessment

For the intent of this paper, formative assessment has three key characteristics, which happen to be the most widely discussed in the literature: 1) purpose, 2) cycle of use, and 3) planned integration with instruction. We will describe each of these three key characteristics in the following paragraphs:

Purpose. Scriven (1967) is widely credited with making the distinction between formative and summative uses of information in the context of program evaluation. Scriven asserts that the former emphasizes the collection, analysis, and reporting of information for program improvement, while the latter has as its purpose rendering judgments on overall program quality or impact.

By analogy, a formative assessment in an education setting is one whose primary purpose is ongoing instructional improvement. In other words, assessment is formative when it is implemented to provide information teachers can use to change the way they teach (e.g., how they present content, group students, give feedback on student work).

In contrast, assessment is summative when its purpose is to provide cumulative judgments about student learning (e.g. student achievement over the course of a school year or other instructional interval). In addition to informing students and their families about their academic progress, summative assessments give stakeholders (e.g., the community, Board of

Education, district, and school personnel) information they can use to decide the effectiveness of a program, an intervention, or a school.

The distinction between improvement and judgment may seem to imply that formative assessment will be classroom-based and summative assessment will be externally imposed; yet, this is not always the case. Some external tests (such as interim or benchmark assessments tailored to curriculum) are designed to be used formatively by providing feedback to teachers who can in turn use them for remediation. Moreover, teachers also assign summative tests in the classroom—such as unit tests and final exams—for the purpose of judging student mastery of large segments of content. In sum, while formative assessment is typically associated with classroom practices and external exams are most often summative, it is possible for both classroom-based and external assessments to serve either purpose.

Thus, the first key characteristic of formative assessment is the way the information is intended to be used. It is important to note that intent alone cannot render an assessment formative; hence, teachers must follow through on this intent with informed action. Black and Wiliam (1988b) incorporate this notion into their definition of formative assessment: "We use the general term *assessment* to refer to all those activities undertaken by teachers—and by their students in assessing themselves—that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes *formative assessment* (i.e., when the evidence is actually used to adapt the teaching to best meet student needs") (p. 2). For Black and Wiliam, action must accompany intent in formative assessment. In our interviews with teachers, we found that certain assessments had a) both formative intent and use, b) formative intent but no apparent formative use, and c) formative use but no clear formative intent.

Cycle of use. Another distinction that is often made to differentiate assessment types relates to how assessments are integrated into the instructional calendar (in particular, the frequency of their administration and the rapidity with which results can be accessed). Assessment for summative purposes typically occurs infrequently, after students have had an opportunity to learn a substantial amount of content. The lag time between instruction and assessment is not a concern if the information is being used primarily for long-term decisions (e.g., as part of a school- or district-level accountability system, such as No Child Left Behind).

In contrast, assessment for formative purposes occurs more frequently, as new content is encountered and insights are needed quickly to inform decisions about how to best teach

the topic at hand. Perie, Marion, and Gong (2007) go so far as to require action within a single lesson, using the phrase ―minute-by-minute assessment" and citing a typical cycle time of five seconds to one hour for formative assessment (p.3). Wiliam (2007) makes a distinction between three cycle lengths for formative assessment that we find to be a useful sorting device. In fact, Wiliam asserts that long-cycle formative assessment is used ―across instructional units, quarters, semesters, years (4 weeks to 1 year)"; medium-cycle formative assessment is used ―between lessons or units (1 day to 4 weeks)"; and short-cycle use occurs ―within a single lesson (5 seconds to 2 hours)". We found examples of assessments of each of these three cycle lengths in our teacher interviews, all with some degree of formative purpose or use. However, we restrict our definition of formative assessment to those assessments of short and medium cycle lengths that are intended to benefit current students and improve ongoing teaching and learning.

Planned integration with instruction. A third feature of formative assessment, which is related to the notion of purpose/intent, is the planned integration of assessment with instruction (i.e., assessment activities designed or selected in order to provide information for instructional improvement and embedded to some degree in instructional activities). Further, opportunities to use this information are an essential element of instructional planning (Black & Wiliam, 1998b; Shepard, 2000; National Council of Teachers of Mathematics, 1991). For example, Shepard et al. (2005) view formative assessment as ―strategies and tools that teachers use as part of everyday instructional routines" in ―recursive assessment processes" essential to improving teaching and learning (p. 277). For Wiliam (2007), formative assessment needs to be ―embedded in the day-to-day life of the classroom" and ―integrated into whatever curriculum scheme is being used" (p. 1091). In other words, assessment is clearly connected to the content currently being taught; in short, it is lesson-related. A final point is that instruction and assessment ideally should form a feedback loop in which instructional activities are regularly adjusted in response to information about student learning that the teacher seeks and obtains during the instruction. Yet, in our teacher interviews we found only a few instances of formative assessment practices that were carefully planned and integrated with instruction in this recursive manner.

For the purposes of this paper, we analyze teachers' assessment practices in terms of a) purpose, b) cycle length, and c) planned integration with instruction; in short, these three features constitute our implicit definition of formative assessment.

Methods

Sampling Districts and Schools

This study was part of a broader assessment project undertaken from 2005 to 2010 by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Please see Appendix A, if you would like a description of that larger research effort. The 2005-2010 study examined the adoption and use of assessment systems characterized as either *interim* or *formative*, which are a wider range of assessments than we consider here. Companion papers will explore interim assessments (Shepard et al., in press) as well as districts' assessment adoption decisions (Davidson and Frohbieter, in press).

Sampling for the broad project, and by extension for this study, involved two parallel searches: First, we looked for assessment systems that fell within the working definitions of interim and formative assessment adopted by the larger project; next, we looked for districts in our geographic areas that implemented such systems. The process was somewhat interactive, as we searched for districts and for assessment systems simultaneously. In the end, we identified three assessment systems that we characterized as formative. We were able to find five districts that were using these systems and who agreed to participate in our study. In the following paragraphs we describe the three formative assessment systems as well as the districts in which we studied their use.

Formative-P. Formative-P represents an early implementation of PowerSource©. This assessment system was designed to illuminate student understanding of a few ~~"b~~big ideas" (e.g., rational number equivalence, the distributive property and equations) that play a central role in middle school mathematics. Researchers at UCLA first analyzed the content of pre-algebra and algebra courses and then identified pivotal ideas that reoccurred throughout the curriculum. The researchers developed a system of assessment exercises and instructional materials that addressed these concepts (which are later described in great detail). In a separate study, CRESST compared student outcomes across math classrooms in grades 6-8 that were randomly assigned to the mature PowerSource© assessment or more traditional assessment-related professional development. As a preliminary finding of this study, Phelan et al. (2009) reported that students in the treatment group significantly outperformed students in the control group on tasks related to the distributive property and that the effect was larger for students with higher pretest scores.

The Formative-P system consists of three elements: 1) sets of short assessments for teachers to use in their classroom four times in the year; 2) materials for three supplemental instructional units; and 3) professional development. Teachers using Formative-P are

expected to teach the three supplemental units on rational number equivalence, ratios, and the distributive property, as well as use the accompanying assessments. For each unit, there are five short Formative-P assessments, called *Checks for Understanding*, which teachers are taught to use and quickly score in order to monitor student understanding. Teachers also receive training on how to teach the underlying mathematical content. They also receive guidance with how to individualize instruction for students whose assessments reveal a failure to master the fundamental concepts tested. The initial professional development session explains the experimental and the theoretical basis of the project. Subsequent sessions follow the sequence of the units and give teachers a chance to look at student assessment data on each unit from their district. There is also a fourth, cumulative set of assessments that are administered at the end of the year.

We selected two districts for our study that CRESST had already recruited to participate in their full experiment (the districts' intent and adoption processes are described in Davidson and Frohbieter, in press).

The first district, Sinclair[1] is an ethnically diverse urban district with more than 200 schools and 100,000 students (over half qualify for free or reduced-price lunch). With a relatively small number of middle schools participating, Sinclair began to implement Formative-P in 2005. Within each participating school, half of the teachers who volunteered to participate were randomly assigned to use Formative-P and to receive accompanying professional development. The balance constituted the comparison group, who received professional development but used traditional assessment. Based on our interviews with three district-level administrators in Sinclair, it appears that the Formative-P effort did not have a large impact on the district's overall math program or existing assessments for middle schools. It seems to have been a limited research activity confined to a particular sample of schools.

Formative-P was also being used in the Adlington district. Adlington is much smaller than Sinclair, with about 20 schools serving roughly 20,000 students. It is also racially diverse, and about 70 percent of students are eligible for free or reduced-price lunch. In Adlington, randomization for Formative-P was conducted at the district level; hence, sixteen middle school teachers volunteered to participate in the experiment, and eight teachers were randomly assigned to the experimental condition. According to the district's Math Specialist, Adlington joined the study (in part) to obtain additional staff development related to mathematics assessment for which it lacked funds. As in Sinclair, district administrators in

[1] Pseudonyms are used for all districts and individual respondents.

Adlington appeared to have little familiarity with Formative-P, and it was not well integrated with the district's existing middle school math assessments. Almost one half of the participating teachers opted not to continue in the Formative-P experiment beyond their initial one-year commitment; this may be related to the district's low level of involvement. Sinclair and Adlington had each developed a relatively comprehensive set of their own assessments with no explicit link to Formative-P, and only a small minority of teachers in each was exposed to or used Formative-P assessments.

Noyce. The Noyce Foundation has developed a system of professional development and math assessments with the goal of developing high-level mathematical skills for students in kindergarten through algebra (Foster & Noyce, 2004; Foster et al., 2007).[2] As of 2008, 35 school districts worked with the Foundation and its partner, the Silicon Valley Mathematics Initiative (SVMI). These districts receive coaching support and professional development for teachers. The districts were also granted access to the annual exam the Foundation created along with related instructional tasks for classroom use.

The signature element of the Noyce math program is the Mathematics Assessment Resource Service (MARS) test, which consists of five items meant to test five core ideas drawn from the curriculum for that grade level. Each of the items is open-ended and involves a multi-part progression from a basic/simple to an advanced/complex understanding of the core idea; thus, it "ramps up" from simple to more complex. The items are intended to assess problem solving, reasoning, communication skills, and the core content ideas. At a minimum, Noyce districts administer a MARS test every March. Teachers are provided with binders that include a bank of MARS tasks, scoring rubrics, and assessment reporting templates (for summarizing student strengths, weaknesses, and the teacher's plans going forward) to use as they wish throughout the year. These tasks are identical in format to those on the annual assessment. Teachers also receive a map linking each task to one or (usually) more of the core content ideas, to assist them with planning. Some districts have also elected to purchase Noyce interim/benchmark assessments, which also consist of five constructed response items with multiple parts. Though these assessments are parallel in structure to the March test, they include a narrower range of content. In many cases, teachers in a school or district also develop monthly unit tests that draw from the task bank to assess content covered during the unit.

[2] For more detail on the Silicon Valley network infrastructure, see: http://www.utdanacenter.org/umln/downloads/sanjose04/sanjose_siliconval.pdf.

The main element of Noyce professional development is the communal teacher scoring of MARS exams. After an initial 90-minute training on a scoring day, teachers rotate exams after a single task is scored. At the end of the day, teachers collectively discuss student responses and how to modify instruction in response. This collective discussion takes place in the presence of content area coaches who have additional training in how to interpret student work on tasks and tie it into future instruction. Teachers gather across districts for scoring after the March exam; they also meet within their districts in smaller groups for scoring the monthly benchmark tests.[3]

In addition to collective scoring, teachers attend a one-week summer institute on algebraic reasoning as well as regular within-district professional development activities. They also meet one-on-one with the district math coach (hired by SVMI). The annual teacher professional development sessions focus on how to redress common points of weakness in student performance on the prior year's March MARS exam. SVMI math coaches lead workshops for teachers consisting of pedagogical coaching and classroom video discussions. There are also sessions for school principals multiple times per year and for emerging leaders and math coaches.[4]

We selected two districts that were working with the Noyce assessments: Prairie and Alta. Each was a small district, with about 15 schools and 7,500 students; Prairie was suburban while Alta was more urban. Both districts were racially diverse with about half of their students eligible for free or reduced-price lunch. Both districts had a multi-year history of working with the Noyce Foundation. After an initial relationship with single schools in each district, the Foundation provided grants to expand training on formative assessment in math to all middle schools. In contrast with the PowerSource© districts, Prairie and Alta have largely oriented their middle school math assessment programs around Noyce products, including a mandatory March summative exam and, in the case of Prairie, interim assessments (in addition to the task bank available to each participating teacher). District staff reported that department leads for every school and nearly all math teachers have participated in Noyce training sessions; interviews with teachers and administrators indicated a relatively high level of familiarity with and frequent use of Noyce materials.

Of these two districts, the Noyce system was most deeply embedded in Prairie's district assessments. Middle school math teachers in the Prairie district had to administer the March

[3] Typically, the benchmarks are scored within districts and the monthlies (when used) are scored at the school level (although the _collective' nature of this is less consistent); depending on opportunities, sometimes teachers got together, sometimes they did not.)

[4] For more detail on the structure of the Noyce network and frequency of meetings, see http://www.utdanacenter.org/umln/downloads/sanjose04/sanjose_siliconval.pdf.

MARS summative exam; moreover, as of 2007, Noyce interim assessments were administered quarterly. Alta also mandates the administration of the Noyce MARS summative exam, but it has developed its own common midterm for grades 4-8 (first piloted in 2007), as well as its own end-of-year exams for the same grades (given in addition to the state summative test). These district exams included guidelines for what should be taught each semester. As the math coach reports: "[Prior to the district exams] there was no consistency in the district about what was taught at grade level. The idea was that we needed sort of a common core."

In addition to the mandated administrations, most teachers in the Prairie and Alta district used the Noyce assessments and tasks regularly; although the frequency varied from teacher to teacher (e.g., from a few times a year to almost weekly). The tasks selected by teachers who used them most often were related to the units currently being covered. Some teachers used the tasks instructionally (such as having students work on them in groups), certain instructors used them as independent quizzes; and yet others used them in a variety of ways for both instruction and assessment.

Freudenthal. Based in the Netherlands (with a chapter in Colorado) the Freudenthal Institute is the largest research and development institute on math education in the world. The Institute began work in the U.S. in 2003 where it focused on implementing a theory of learning and teaching math called Realistic Mathematics Education (which claims that students should have a "guided" opportunity to "re-invent" math by doing it). To enact the theory, the Institute supports teachers' development of classroom assessment and curricula; furthermore, the Institute helps schools and districts build leadership capacity through professional communities.

Unlike Noyce and Formative-P, the Freudenthal professional development does not include a system of embedded assessment materials. Instead, the Freudenthal Institute has developed a framework for assessment called the assessment pyramid, which classifies three levels of mathematical competencies: 1) reproduction (i.e., recall), 2) integrating mathematical tools, and 3) analyzing. The Institute recommends that assessments include questions that cover all three levels of thinking. The Institute also provides an online series of algebra modules for teachers' use that provide sample tasks classified into the three aforementioned levels of competencies. In the Jackson School District, the Freudenthal Institute offered a two-week summer institute for middle-school math teachers, six two-hour school learning community meetings during the year, and four full-day, district-wide professional development days during the year. These meetings focused on specific strategies for teaching core concepts as well as ways to assess student understanding of those core

mathematical concepts. The Freudenthal training encouraged frequent, lesson-focused assessments that were much narrower in scope than Formative-P and Noyce.

Jackson was a suburban district with about 50 schools and 40,000 students. About three quarters of the students in this district were white, and only about 15 percent were eligible for free or reduced-price lunch. In contrast to the standardized assessments in place in the Noyce and Formative-P districts, Jackson had a decentralized system of math assessments for middle school. It did not impose strict guidelines on its math teachers regarding assessment schedules and types; nor did it mandate any middle school math test aside from the statewide summative exam. The Chief of Planning and Assessment stated: "I know that teachers have a whole mix of practices ranging from using the unit assessments and various assessments with the text series, to using the online assessments that come with some of these, and then of course, the substantial amount of teacher-generated assessments."

Freudenthal training was one of several professional development options available on a voluntary basis to math teachers in the district, and it involved fewer than half of the district's middle schools. The district administrators we interviewed mentioned Freudenthal training in the context of professional development options for middle school math that focus on instructional strategies. The teachers we interviewed discussed the Freudenthal professional development to varying degrees (with some describing extensive participation and others not mentioning it at all). The teachers at one of the two schools we studied in the Jackson district generally discussed this program more than instructors at the other school.

Sampling Teachers

The teachers who participated in this study represent a convenience sample drawn from all teachers who were using the identified formative assessment system in each of the selected districts. Specific sampling methods varied to accommodate differences in district sizes and prevalence of use of the targeted assessment. In Jackson, district administrators identified two schools that were participating in the Freudenthal professional development; the principals of these schools each selected three teachers to participate. In the Noyce and Formative-P districts, we invited all teachers using the assessments to participate in the study; we also included all teachers who volunteered in our interviews. Teachers received a small honorarium for agreeing to participate in two interviews.

Data Collection

We used a two-part interview process to try to learn as much as possible about each teacher's assessment practices. The first part of the interview was conducted by telephone and teachers were asked about their uses of assessment. For the second part of the interview

(which was conducted in person at each teacher's school), we requested that teachers bring samples of some of the assessments that they used. During the in-person session, we asked further questions about assessment use. We also requested teachers to describe specific information they gleaned from their assessment examples as well as the ways in which they used this information. Lastly, we asked to keep copies of these samples (with the names of students removed). Each interview was audio-recorded and transcribed; also, the assessment artifacts we collected from the teachers were digitized. Hence, electronic versions of all interviews and artifacts were created, which could be shared and discussed among the geographically dispersed research team.

Data Analysis

When approximately half of the interviews had been conducted, we started our analysis by reading through the transcripts in an unstructured way, and then by summarizing teacher cases individually and in pairs. We met periodically to flesh out our initial observations and discuss possible ways of coding the data. We arrived at an outline, which was based on assessment types and features, information gained, and uses of the information. We continued to add potential codes under these headings until we arrived at a draft code list. Using the NVivo software program, we began a process of refining our code list by coding cases in groups (first the whole research team and then in pairs) and meeting to reconcile our coding.

After we arrived at a final code list, we continued the process of pair coding and reconciliation, and mixing pairs as a means of providing consistency across the research team. To reconcile, a pair would exchange coded transcripts and discuss segments for which they disagreed, deciding on the most appropriate coding for each segment. This process produced a master NVivo file with all transcripts coded. Once coding was completed, we divided up into two research groups to study districts that were using interim and formative assessment systems separately. Two districts, Adlington and Sinclair, were included in both subsets of the data. Each research group used the master NVivo file to generate reports based on codes and groups of codes using Boolean logic; these reports formed the basis for each research group's detailed analysis.

Results

We will present the results of the study in five sections. First, we provide a brief overview of the range of assessment activities described by teachers in the study, comparing responses from teachers using each of the three sampled assessment systems. Second, we tackle our first research question and examine the kinds of information teachers learned from

the formative assessment they administered. Third, we address our second research question and explore the ways teachers used the information they obtained from the assessments. Fourth, we describe a pattern of assessment-related practices that we characterize as ―true" formative assessment because it embodies the three elements of our definition. Finally, we target our third research question and compare the teachers' reports of information and use across the three assessment systems, examining the relationship between the type of formative assessment (i.e., the three systems we examined), the information teachers obtained, and the ways they used this information.

Range of Assessment Activities

When asked about the range of activities they used to learn what their students knew, teachers mentioned many types of assessments; the topics they discussed varied across the three formative assessment systems we studied. The teachers using the Noyce system talked mostly about the Noyce assessments and tasks, and less about other activities they might have been using for assessment in their classrooms. The teachers using Formative-P tended to speak about the Formative-P materials as well as other types of classroom assessments (such as quizzes and tests that came with the text series they were using). As the Freudenthal professional development does not introduce new assessments or instructional materials, the teachers who had undergone this training rarely referred to Freudenthal when discussing the assessments they used; rather, they talked about a variety of activities used in their classrooms for assessment. Although the districts were each selected because they were using one of the assessment systems we were interested in, we chose to explore the full range of assessment activities the teachers described rather than limit our analysis of formative assessment to activities directly associated with the targeted systems. This analysis provided information about a wide range of topics; yet, in the following sections we will focus on answers to the three research questions presented previously, beginning with what information teachers gleaned about students from the various assessments they used for mathematics.

Research Question 1: Information Teachers Gained From Assessments

As would be expected, most of the information teachers reported that they had obtained from assessments was related to their students' *mathematical skills and knowledge*. Many teachers described learning information about their students that was not directly related to specific math content—including issues of motivation, study skills, as well as students' abilities and willingness to convey mathematical information in writing. Though interesting in other ways, this type of information does not pertain directly to formative assessment;

thus, we limit our discussion to the information teachers gained that was related to the math content taught. Their remarks about what they learned ranged from simple binary statements (e.g., "I learn whether they get it or not") to nuanced observations about students' understandings of detailed aspects of mathematics. In the next section, we will describe these classifications further, and then present examples reported by teachers in our study.

The range of "nuance" in the information teachers described. We classify the math content-related information (that teachers had reported they had obtained) according to three levels of *nuance*. At the low end of this spectrum are *binary* remarks that refer simply to *whether or not* students know or can do something (i.e., the presence vs. absence of knowledge). Teachers used phrases such as "Do they get it" or "They can't multiply by ten" in discussions of information at this level. *Moderately nuanced* examples of information include something about the nature of the students' knowledge or lack thereof. Two of the most common characteristics of knowledge teachers discussed at this level were (a) depth of understanding, and (b) whether errors were procedural or conceptual in nature. The most *highly nuanced* examples of information teachers described to us involved detailed explorations of students' thinking, such as examining where a solution process broke down and why, or describing a partially-correct conception that produced right answers only in certain cases.

Binary judgments about mathematical skills and knowledge. Many teachers began their discussions of the information they gained from assessments with *whether-or-not* remarks, as in Michael's explanation of a strategy for a weekly assessment:

> …The formative tests frequently happen every Friday. So we'll do the instruction Monday/Tuesday and practice it throughout the week and then see *whether or not* they learned it on Friday.

Though he uses the term *formative*, Michael's description suggests a more summative purpose for the assessment, as it is deliberately given at the end of a period of instruction. In many cases, however, teachers' descriptions of *whether-or-not* information and its uses do apply to several aspects of formative assessment.

Most often, teachers described information in a binary way when the information was sought explicitly to ascertain students' readiness for instruction (whether a single lesson, unit, or course). Several teachers discussed warm-up exercises given at the beginning of a lesson to obtain information about students' readiness to proceed. Joan, for example, explains that warm-ups give her "a quick assessment of whether students are really grasping the concepts

or not." Joan also went on to describe instructional actions that she would take during the lesson if none of her students answered the warm-up problems correctly.

Robert discusses the information about his students' mastery of the previous day's lesson provided by warm-up exercises:

> …Assuming that the kids got the lesson, and everything went fine, 90% of them scored. Now if they don't get the lesson, and it's a newer topic, like finding area of a circle, that was one where instead of the average classroom getting 90%, the average classroom would get 60%. And then I would say—Oh, I've got to re-teach that lesson.

Robert views the class average on the warm-up as a way of determining whether his students "get the lesson" (e.g., 90% average score) or "don't get the lesson" (e.g., 60% average score), and he bases significant instructional decisions on this information, as will be discussed further in the *Teachers' Uses of Information* section of this report.

Several teachers reported using some form of pre-assessment at the beginning of an instructional unit in order to learn whether or not students had already mastered the topics or subtopics that would be covered. Leah, for example, uses a pre-assessment to gauge a class's prior knowledge of a unit on fractions, decimals, and percents:

> … The way that I end up using [the pre-assessment] is that if there is one question that everybody gets right – I don't have to cover that. If there's one question that everybody is getting wrong, I have to cover that. The stuff in the middle is probably gonna get covered even though some kids know it, because there's such a differing, coming in from all different elementaries.

Leah does not indicate that she explores her students' thinking beyond a judgment of whether or not they "know" each question, which she appears to consider adequate for the purpose of this assessment. Several teachers also reported giving a diagnostic test at the beginning of the year to determine whether their students had the requisite knowledge for a course, in some cases implying they made instructional use of the information they obtained as they taught various topics throughout the year.

In the previous examples, teachers used assessments to determine whether or not a group of students had a specific skill or set of skills. Despite the low level of nuance in this type of information, key features of formative assessment are evident. Teachers used these assessments intentionally to obtain information on which to base instructional decisions, which indicates a formative purpose. Assessments, such as Leah's warm-up exercises, are clearly integrated with instruction in a planned way. Many of the examples of binary-type information were also associated with short-cycle (e.g., warm-ups) and medium-cycle (e.g.,

unit pre-test) assessments, while beginning-of-year diagnostic tests are an example of long-cycle assessment (Wiliam, 2007).

While all of the teachers in our sample described some of the information they obtained from assessment in this *whether-or-not* manner, each of them also provided at least a few examples of more nuanced information.

Moderately nuanced information addressing the nature of students' knowledge. Over the course of the two-interview series, often in response to interviewer probes, most teachers gave increasingly nuanced examples of information they had obtained. In addition to speaking about whether students had learned, teachers described learning information about how, how well, or how deeply they had learned. Here, Don describes obtaining information about the nature of his students' understanding from the Noyce assessments:

> And the information we get is very interesting. We get information about how well they've learned the concept, whether they understand the concept on a superficial level or whether they really get under it and understand it. We get how they are able to learn the skills necessary to accomplish the task and whether they have learned the mechanics necessary.

Don is saying more *about* his students' knowledge than simply whether it is present or absent. In addition to learning whether they have the necessary "mechanics" to accomplish a task, he implies that he obtains information about the depth of their understanding regarding the concept underlying the task.

Some teachers reported gaining information about varying *degrees* of mastery and various instructional needs among their students. For instance, Jeff shared:

> …[The quiz] tells me that they're in really different places, and it's hard to address this all, and I need this person to sit in a small group with other people who have the same understanding… and this person needs some enrichment type of thing.

Teachers also reported learning about students' degrees of mastery of multiple math topics, concepts, or procedures from a single assessment. Patrick, for example, obtains information on students' levels of mastery of a variety of topics from the Formative-P assessment; he finds this information helpful as he reflects on his teaching:

> …I can get the information that shows me which of those areas my kids were doing best in, and which worse, and compared to the other ones. So that is particularly helpful. For example I've found that my kids do real well on the algebra and functions, because I'm probably emphasizing algebra and functions—I'm an algebra teacher for most of my 20-some years of teaching. So I do that, and the measurement/geometry is something I put off, we're going over that now, they didn't have any of this.

Nancy describes the information she obtains from the Noyce assessments in a similar way. She discusses broad areas of ―strength" and ―weakness" for her students as revealed by the quarterly assessments, and information about more specific subtopics in the monthly assessments:

> And so I would tend to see broad sections that they needed more reinforcement for, and then in the quarterlies it was big-picture stuff, like the geometry section they don't quite understand or haven't applied enough or haven't had enough experience with. But in the month-to-month kind, you would see more, like specific skills, like if geometry's a general weakness, then specifically relating area and perimeter as a ratio relationship would be like a specific thing that they weren't able to deal with yet.

Many teachers referred to specific pieces of mathematics when describing information about students' knowledge. For example, Joan explains that a unit test revealed that her students' errors in using the quadratic formula were computational rather than conceptual in nature:

> So it really tells me whether they're understanding the big picture... The most recent was the quadratic formula. Were they able to apply the quadratic formula in solving quadratic equations? And most of the students were. It wasn't the fact that they didn't know how to apply the formula; their mistakes would come in from computational problems. So it just gives me immediate feedback then: did they understand the concept?

Though she does not describe the nature or source of her students' computational errors, Joan is distinguishing this type of error from a lack of conceptual understanding, rather than simply determining whether or not they know the quadratic formula.

An additional kind of information teachers reported obtaining, with respect to the nature of their students' knowledge, was common errors made by students. Mary, for example, explains that when adding and subtracting fractions, her students ―just love to add them together even though they've got different denominators." Mary had to ―stop and say, you've got to have the same denominator in order for you to add or subtract fractions, and they just do all kinds of different things." While this information is somewhat nuanced (in that it points to a specific step in a procedure that is a source of frequent errors), Mary stops short of probing for the reason behind this error in her students' conceptions of fractions.

In the examples provided in this section, teachers discuss information about students' mathematical skills and understanding that goes beyond a binary level to reveal varying degrees of mastery, relative mastery of topics, types of mastery (e.g., procedural vs. conceptual), and common errors.

Highly nuanced insights into student understanding. The most nuanced examples of information teachers described gathering from assessments involved *explorations of students' thinking*, such as examining where a solution process broke down and why, or describing a partially-correct conception that produced right answers only in certain cases. These discussions usually occurred when teachers were examining samples of student work with the interviewer; several examples of these artifacts are included here. In addition to written tasks used for assessment, many teachers described informal assessments they conduct by circulating among their students as they work independently or in small groups. Here, Don describes the nuanced information he looks for in such informal observations:

> I try to understand the flaw or the deficiency in their thinking. What is this child not seeing or not understanding that's preventing them from succeeding in this? I try to pinpoint the one thing that they don't get. And I try to find a way to talk about that misunderstanding or that thing or that flaw they don't get, to find a way for us to talk about it.

Don is clearly seeking information that will allow him to address specific aspects of the student's thinking and help him move forward with his understanding.

A common type of nuanced insight teachers described was finding out which piece of a specific mathematical process was the source of students' errors and *why*. Patrick, for example, uses a quiz to learn about detailed aspects of students' skills and knowledge with respect to the division of fractions:

> I'm looking, for this particular thing, where are they getting stuck? Can they do the first step, which for example, on a division of fractions [problem], did they take a mixed number and change it into an improper fraction…? All right, if they can't do that, then that's one problem. Could they take the second fraction in the division problem and know that they had to flip it over to its inverse or reciprocal? And if they didn't do that, that's a different problem, and maybe it could be fixed right away, but there's some kids that for some reason don't know why it's the second fraction, and might do the first fraction, and think, as long as they did one, it should be all right. So did they get that setup? And then for again, fractions, I'm looking if they've gotten that far, did they go ahead and cross reduce first? Or did they multiply and then try and reduce, which is a much harder way to do it, so why didn't they cross reduce?

Patrick is exploring detailed aspects of students' computational processes, attending also to their understanding of the algorithm's logic (they ~~–~~don't know why it's the second fraction") and the reasoning behind their decisions as to whether and when to reduce. This type of information provides a strong basis for providing targeted feedback that addresses particular misconceptions and sources of error.

Jeff describes constructing formative quizzes in a way that allows him to sort out specific aspects of his students‘ skills and understanding. Using the topic of fractions as a hypothetical example, he discusses this assessment-building technique:

> Well, each problem I put on an assessment is specifically designed for me to learn if they understand how to do something or not. So there might be something very basic, like adding fractions, and then the next question might have them add fractions but then also require them to simplify it. There might be then, also, a problem where it's more of a problem solving, so with the first question, I get to see if they can just do the algorithm, so I know, can they even do the algorithm, so it gives me some data. Then is the issue later, did they get the problem-solving one wrong, so then I at least know where to start with that student… So the idea is each of the five questions on a formative is very specific to a scale that I'd like them to have... I'm looking for particular mistakes and looking for particular answers.

This creation and use of assessment is planned in seeking evidence of specific misunderstandings to be addressed. An example of a set of Jeff‘s linked quiz items is shown in Figure 1. He describes the relationship between items 2 and 3:

> Here (question 2), the idea is they need to divide up the pet food amongst the three boa constrictors so it's going to be [seven tenths] divided by three. In here (question 3), I was sneaky. I used the same two numbers, but it's reversed. Instead of seven-tenths divided by three, it's three divided by seven-tenths. And to recognize that, and then it teaches me who's there. And it teaches them. The other thing is it teaches them, ―Oh, I really need to pay attention here.” Because in the past, a lot of the kids could go, ―Oh, I know I have to divide the small number into the big number,” and they didn't have to really think about clues in word problems.

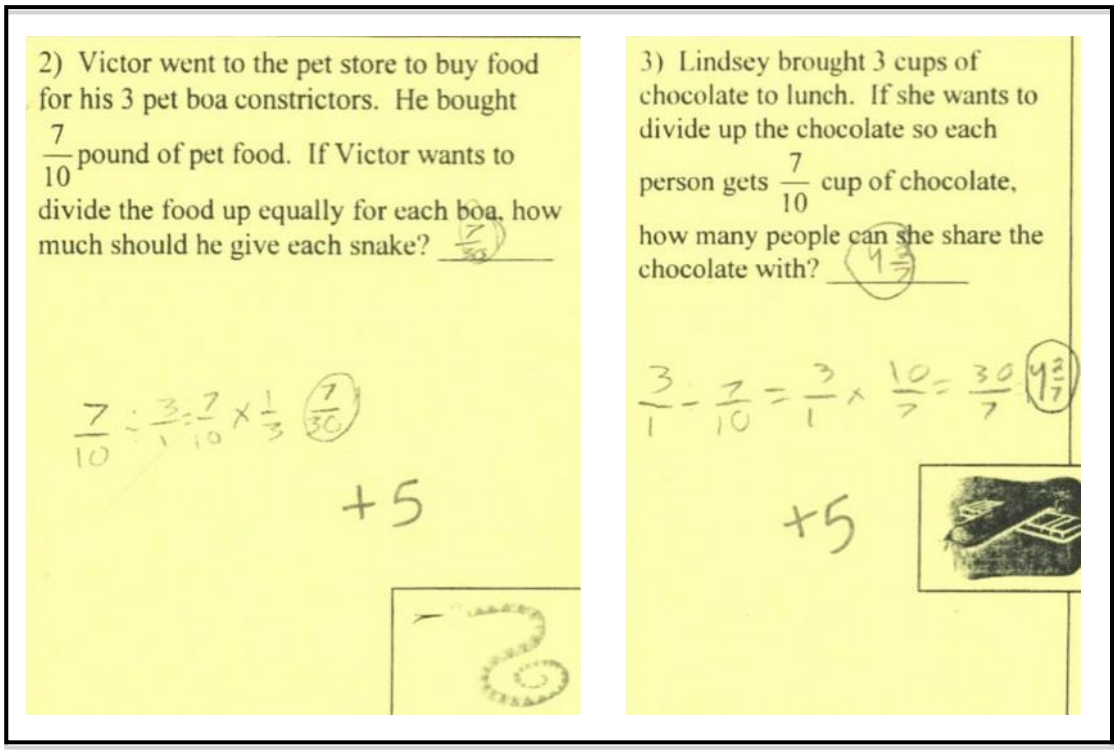


Figure 1. Linked items (2 and 3) on a formative quiz.

It is interesting to note that in addition to seeking information, Jeff appears to use this assessment as a teaching tool. He intends the pair of problems to force students to ―pay attention" to the meaning of the numbers rather than assuming which is the divisor and which is the dividend based on their experience dividing smaller numbers into larger numbers. This is one of a small handful of examples in which teachers explicitly referred to a task being used for both assessment and instruction (though Jeff appears to have missed an opportunity to provide guidance on the appropriateness of expressing a fractional number of people).

Like Jeff, many of the teachers described nuanced information they had gained about individual students, typically analyzing a student's incomplete understanding of and/or source of error on one or more of the items on an assessment. Here, Ann discusses a student's partially correct response on a Noyce task (see Figure 2 on the following page):

> And then this was an example of a typical student… They were just looking at enlarging one of the sides rather than the border. So when they interpreted border, it was just one of the sides. So here it was like eight, [counting squares] one, two, three, four, five, six, seven, eight. And they knew that it was enlarged three times. All they did was enlarge, I think, the base of the rectangle. So they knew this idea of enlarging and that you could just multiply it by three. But they didn't understand the concept that it was the entire border…

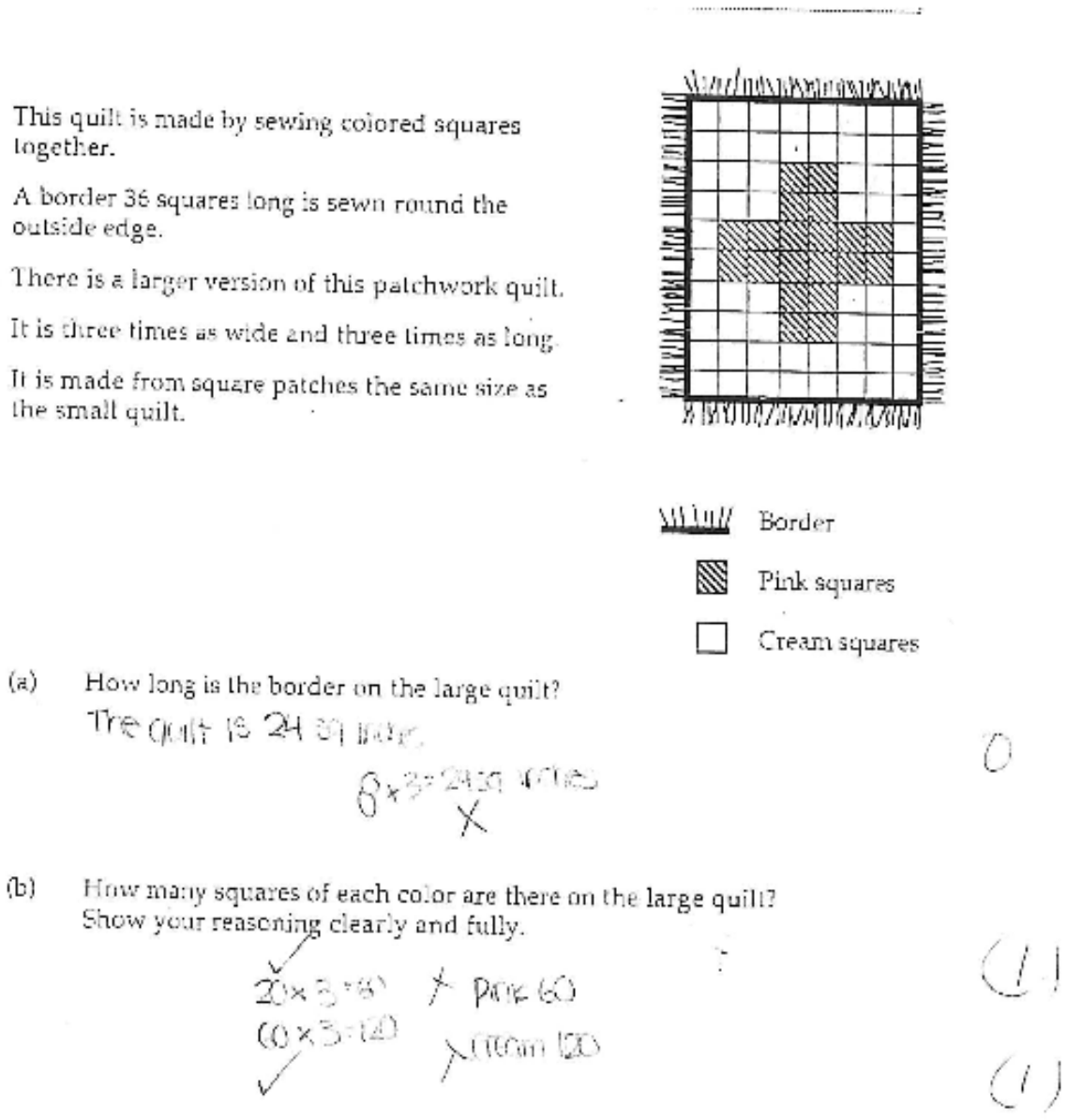


Figure 2. Partially-correct student work on a Noyce task.

Ann recognizes the student's understanding of the relationship between ~~–e~~nlarging" and multiplying, even though the student stopped short of correctly applying this relationship to the border and number of squares in the quilt. Ann's analysis of the student's partial understanding of ~~–e~~nlargement" provides her with a stronger basis for providing feedback and/or tailored instruction than she would have had from simply noting the incorrect response. She later explains that the open-ended nature of the Noyce tasks is conducive to illuminating students' thinking and helping her learn *how* they arrive at their responses:

> I definitely like it better than, like a multiple-choice test, because I'm able to see their thinking a lot more, at least I feel that I'm able to see their thinking a lot more. Like when it asks them to explain their reasoning and the students actually articulate how they got their answer, I'm able to see that. I feel like knowing how they got an answer… helps me a lot more than the answer itself, if that makes sense.

In an example of an insight related to a subset of students within a class, Leah perceives a specific misconception during an informal assessment on subtracting integers using a chip board, which is apparently a graphical device:

> I'd say there was almost half the class that was actually just—they decided that [since] subtraction and addition are opposites and negative numbers are opposites you must just add instead of subtract. Like they had created this idea in their heads which I thought was fascinating. And actually it is true in some situations—it just wasn't true in all of them. And I got that from looking at that chip board. And I was able to address that.

This example contrasts with Mary's fraction addition example in the previous section, in that Leah discovers the idea that led her students to make mistakes in some cases. This insight provided Leah with a strong basis for ~~a~~ddressing" this error with her students, in that she could build on their partially-correct conceptions.

In the previous examples, the teachers obtain highly nuanced and detailed insights into the thinking behind their students' mathematical work. In seeking to understand their thinking—regardless of the correctness of their answers—they learn information that can help them assist their students in moving forward in their understanding of mathematics. We believe these highly nuanced insights represent the type of understanding that can best help teachers to ~~m~~odify teaching and learning activities" to ~~b~~est meet student needs" (Dylan & Wiliam, 1998b, p. 2). By exploring their students' thinking, teachers can uncover ~~a~~ctive student conceptions that may hinder learning," allowing them to effectively tailor their instruction in the formative assessment process (Wiliam, 2007, p. 1070). Highly nuanced information is discussed further in the following section on ~~t~~rue" formative assessment, along with a description of how that information is put to use by teachers to plan and adjust instruction.

Research Question 2: Teachers' Uses of Information From Assessments

This section considers the range of actions teachers take to plan or adjust instruction in response to the information they gather from their classroom assessments. Teachers reported using assessments in ways that varied considerably in terms of how responsive they were to the information gathered from the assessments; this section describes examples in three categories from least to most responsive, ranging from no action at all to revising lesson

plans and re-teaching the topic. Most teachers did report taking at least some action in response to assessment information. The most common actions were re-teaching prior lessons, reviewing and/or correcting completed assessments with the class, and assigning additional problem sets.

Each of these three types of actions could be done with different degrees of sensitivity and customization, which are the bases on which we classify them as either moderately- or highly-responsive. For example, re-teaching strategies ranged from repeating all or parts of the original lesson with little modification to making substantive adjustments in response to particular aspects of student difficulty. Similarly, strategies teachers employed to review completed assessments ranged from the simple reporting of incorrect answers to the analysis and correction of errors by students and teachers together. Finally, assessment-related strategies for assigning additional work ranged from class-wide homework assignments of problem sets focusing on the most common areas of difficulty to more sensitive and customized structured opportunities for guided practice on problem sets and, on occasion, problem sets tailored to individual or small groups of students.

It is important to note that the reported uses of assessment information varied both between and within teachers (i.e., not only did the range of practices differ from one teacher to another, but each teacher described different kinds of uses in his or her own practice). However, the data did suggest patterns of action that were associated with the different focal assessment systems we studied. Some of these patterns are explored in the following *Comparisons* section.

Least responsive actions: Move on. As stated previously, the least responsive action one can take in response to assessment information is no action at all. In fact, a few teachers reported that sometimes—even when assessment results convinced them that students did not learn the content of a recent lesson—they simply moved on to the next topic. Teachers gave a variety of reasons for deciding not to address deficiencies but to take up the next topic, including a rigorous district pacing schedule that left no time to revisit material, the breadth of material they were required to cover during the term, the demands of a large class size, and the pressure to focus on content most heavily represented on district tests.

Moderately responsive actions. The most common action teachers reported taking in response to student difficulty with the material was to provide a review of all or part of the original lesson with little revision beyond the use of simplified example problems. This type of re-teaching typically involved breaking down mathematical procedures outlined in the original lesson into more basic steps, for example, Robert revealed:

> We've already gone through the lesson in the book, so I go through the publisher, and I say, all right, take the exact same lesson, then simplify it down a little bit, fewer homework problems on it, and it shows them how to do it a slightly different way, or step-by-step way. And if for whatever reason I didn‘t cover it, it gives them one more chance at sinking in.

Teachers often reported that these class-wide reviews focused on problems students had missed on completed assessments or classwork. While a few teachers reported simply providing students with correct answers and little explanation, most teachers reported modeling a correct way to solve missed problems. In the next example, Janet reports using the review of completed assessment tasks as an opportunity to improve students‘ procedural fluency in solving similar problems:

> What I do is, after I give them the chapter test, which I try to give to them immediately, I ask them—okay, I go over one, is there anybody that wants me to go over one? Two, three… so I kind of do a roll call and if any student has a problem, that's when I'm up there with the transparency and I don't break it down like this, I just write the problem myself and I show them, "Okay, this is step one, step two, step three, step four…" and so forth. But I do review it with them.

Some teachers also reported incorporating students‘ solution attempts into the review of missed problems. This typically involved a show-and-tell approach, in which a correct student solution was used as a model for the class to follow. Patricia shared:

> Here's the question where people had difficulty and here was what the difficulty was, so as a class you can look at a question and you can see, okay here's what they didn't get... So you can explain that this is where most people got it wrong and here's what it's about. And you can do it by showing students' work who did it really well. I like to show that instead of just going blah, blah, blah. They didn't get it when I blah, blah, blahed before.

Finally, teachers often reported using information from assessments to guide how they assigned additional work. Teachers typically reported that they assigned class-level problem sets that were responsive to what they perceived from an assessment to be the modal area of student difficulty. While warm-ups are often discussed in the literature as a formative assessment strategy, several teachers who discussed warm-ups emphasized their use as an opportunity to give students more practice with particular concepts or types of problems. In fact, the teachers in our study discussed warm-ups as a *response* to information from assessments as much as a *source* of information about student learning. A typical example of this responsive variety of the warm-up is discussed above in the *Information* section. One teacher in our study, Robert, includes similar problems on future warm-ups when the class average on the current warm-up falls below a certain threshold (see page 14 of this report).

In addition to these relatively short-cycle uses of assessment to assign additional practice as needed, some teachers reported taking advantage of a spiraling curriculum (in which key concepts are repeated throughout the year at increasing levels of complexity) to address areas of student difficulty in the course of future units. In the following example, Nancy discusses her strategic use of the spiraling curriculum as well as a ―spot review" in which she periodically inserts previously covered material into future units:

> We do re-teaching, reinforcement... And try, even as we go on to new topics… with the scope and sequence it'll spiral and overlap. So we try and hit that kind of problem [with which students had difficulty] a little more frequently than some of the others. So I would be practicing a little more, like for example, ratios in relating one thing to the other with, like length and width, or area and length, perimeter and length. And so even when we're talking into like, systems of equations, there will be, like one or two problems thrown in for homework or something. And then just spot review things as we went along so that they would not forget, or keep it in mind while they're learning new things; not lose it.

In this case, the information Nancy gathers from assessments is regularly used to adjust instruction; yet, those adjustments occur in a longer cycle, across units, rather than during the initial teaching of the topic.

Highly responsive actions. Occasionally, teachers made use of information about student learning to plan and adjust instruction in ways that were highly responsive to the perceived learning needs of students. Here, we discuss highly responsive uses of information in the forms of re-teaching, grouping, and the use of self- and/or peer assessments.

When describing re-teaching, a few teachers reported that they not only revised the original lesson to target areas of difficulty, but they also tailored the mode of instruction in an effort to build on perceived student strengths. In the following example, Janet discusses using ―item analysis" to identify areas of difficulty and adjust her mode of instruction to tap into perceived student learning proclivities. For her, this involved introducing an alternative approach to solving similar problems:

> What I do is item analysis, which… gives you more of a—this percentage of students received this many errors, and from that I review with the students what they need to do. And I noticed, even in the last… benchmark—for example I noticed that I had a large majority of students not do well with the word problem(s). So I had to use a different strategy. I used the… it's called the "four plane," where… they break down the equation and then the solution. Because I noticed that a lot of my students are visual kinesthetic learners, and when they just see the word problem or when they hear the word problem, it seems not to—they get scared. But I showed them how to break it down.

This example offers a compelling contrast to the *simplify and repeat* approach, discussed previously, which typically emphasized fidelity to the originally-taught problem-solving technique. Here, the teacher used information from an assessment, along with her knowledge of her students‘ strengths, to plan instruction that was responsive to their learning needs.

Teachers often reported placing students in small groups to work on problems that were assigned in response to information from assessments, in some cases using the assessment information to form the groups. In these instances, grouping was either mixed-ability (i.e., higher-performing students were matched with lower-performing students) or like-ability (i.e., students with similar performance levels were grouped together). Teachers described mixed-ability groups more frequently. In the next example, Robert discusses his general approach to class work, which is intended to provide lower-performing students with an opportunity to learn from their higher-performing peers:

> What I will do in the classroom setup is I usually like to spread out, and an advanced student in each one of the clusters of my table groups here. And I have one advanced in this four, one advanced in this four, advanced in each group of four. And then that way there's somebody in each little mini-group that can assist, and that's where I look originally, and I pretty much take that year-round, and try to keep a top student in each cluster.

In contrast to this fairly common practice of assigning mixed-ability groups in a whole class, a few teachers reported like-ability grouping strategies intended to differentiate instruction for individual or small groups of students. In the example that follows, Daniel reports using information from quizzes to target groups of students for additional support:

> With [quizzes] I can also pinpoint certain students or certain groups of students that may need extra help with a certain math concept. For example, finding a measurement of an angle when you‘re given the other two measurements of a triangle. So the quiz helps me understand or know that certain students understand the math concepts and then those that don‘t, I can set them aside and keep them after school or do a side mini-lesson with them so they can have a little bit more time, a little bit more of my attention to help them understand the math concept.

Sometimes, teachers reported including self-assessment opportunities in the review of previous work. As discussed in the previous section which described moderately responsive actions, our interviews suggested that this practice often involved simply asking students to correct an assignment (usually a quiz or a test) and then giving them an opportunity to re-take it. However, some teachers reported using self-assessment and/or peer assessment in ways that went beyond simple test corrections. In these instances, students were tasked to consider,

and revise if necessary, their own mathematical reasoning and/or that of their peers. In the examples that follow, teachers discuss pushing students to analyze and revise errors and develop explanations, or justifications, of their own mathematical thinking (and sometimes that of their peers). In these instances, an assessment that may have had a summative purpose was also used formatively, as students received (and sometime participated in constructing) timely feedback on their performance and used this feedback to improve their understanding of the material.

As a first example of this richer mode of test corrections, Jeff describes how self-assessment operates in his classroom, emphasizing that the point of the task is for students to think mathematically and to demonstrate their thinking:

> The other piece of formative assessment I use is they correct all of their assessments: quizzes and tests. And they are supposed to take the problems they got wrong and redo them and… that's not independent, so they can work with me, anybody they want, to get those problems right... Typically, I pick for them or sometimes I let them choose from a few questions on the quiz or test where they have to break it down step by step—the question—and describe why they do each step… So [with corrections], they break the problem down by operation. "I added because of this, and I divided because of this"… Breaking the steps down and really getting it so that they can describe their thinking. It makes them really think.

Jeff and Leah, both Freudenthal teachers, referred to this self-assessment strategy as *meaningful corrections*. In this and other examples in this section, teachers used information from assessments in ways that were highly responsive in some way—by taking advantage of students' strengths to provide alternate teaching strategies, providing opportunities for peer assistance, working with small groups of students on specific areas of difficulty, as well as fostering self-assessment and metacognition. In doing so, they used the information at their disposal in a manner that was sensitive to students' performance and customized to students' learning needs.

Practices That Combine Information and Use in "True" Formative Assessment

Unfortunately, it was rare to encounter instances in which practices embodied all the ideal characteristics of formative assessment (i.e., assessment with instructional improvement as its purpose), which occurred frequently and were related to content currently being taught, as well as were integrated thoughtfully with instruction. However, instances of such improvement-oriented, regular, and planned integrated assessments were not entirely absent from our data; in fact, a few teachers described instances of assessment that met all three criteria. In addition, these instances tended to involve highly nuanced information about

lesson-related content knowledge that was used for highly responsive instructional adjustments. In the following section we describe a few of the cases in which assessment and instruction operated in a close feedback loop, which embodies our definition of formative assessment.

Purpose. The first characteristic of formative assessment we identified in the introduction was that it be intended—and then actually used—to improve teaching and learning. We found many instances of assessment that were intended and used for these purposes, most notably teachers' informal observations of students and some of the uses of warm-up activities. Some teachers also described quizzes that were given for the purpose of adjusting instruction, though these typically also served summative purposes (insofar as they constituted part of students' grades). We found that even assessments that teachers intended mainly for summative purposes were also used formatively on occasion (e.g., when students reworked a unit test and explained how they had corrected their thinking). Among all of the assessments teachers described to us, we found a wide range of practices that met the criterion of formative *purpose* to at least some degree.

Cycle of use. In addition, practices that could be characterized as "true" formative assessment involved the use of assessment information for either immediate, on-the-spot adjustments to instruction, or for medium-cycle planning related to the current instructional topic (e.g., the following day or week). In the next example, Ann discusses the micro-adjustments she makes based on what she gleans from monitoring students as they work:

> So with, I guess, just walking around, it just gives me a general feel for, like, if I need to stop the class and do a quick check-in. And what I mean by that is where I stop the entire class and I say, "Okay. I'm seeing a lot of people do this, and we need to talk about how you're all solving No. 3," or something. So when I walk around, it's just kind of for re-teaching in the moment for that day.

Patricia discusses daily questioning techniques designed to elicit information that she can use in a slightly longer cycle, for the next day's lesson. She explains:

> It's the same kind of information you get from that more formal [assessment] maybe that you give every two weeks. But… it's closer to the time, but you're teaching it so you know really how to readjust the next day.

In addition to making these types of short-term adjustments to instruction, some teachers described how information from a medium-cycle assessment guides their planning for the current instructional unit, as in Ann's description of her use of a Noyce task:

> See, I plan on a weekly basis. So, based on what I see them do. Like after seeing this… I know I need to focus more on what happens to the perimeter and area when something increases, or when something gets enlarged or when the dimensions get enlarged. So I would probably do a warm-up or so, or a couple warm-ups for the next two weeks that deal with this… I use it more as, okay, this is what I'm gonna need to re-teach, or this is what my next week's homework's gonna focus on…

Ann indicates that this adjustment to her teaching of a geometry topic is typical of her use of information from the Noyce tasks to plan instruction on a weekly basis.

As discussed in the introduction, most of the conceptions of formative assessment in current research literature are limited to short- and medium-cycle assessments such as these, which benefit the students who are being assessed. However, we found instances of assessments also being used in a long cycle length, such as from one school year to the next, aimed to improve instruction. Here, Ann continues the previous discussion of how she will use the information from the Noyce task:

> …I think with this one, though, I was just so disheartened by it that it was more of like, I'm gonna change the whole unit next year. And when I teach scale factor this summer, these are the notes that I know I can look at it, and I'm gonna rearrange it or spend more time giving them more opportunities.

While such long-cycle uses do not fit within our definition of formative assessment, it is worth noting that we found several instances in which assessment was used for instructional improvement even across school years.

Planned integration with instruction. Finally, in cases where assessment and instruction were tightly integrated, the distinction between the two was less clear. That is to say, assessment strategies often served educative purposes; student progress on instructional tasks was closely monitored with an eye toward on-the-spot adjustments. One example of assessment regularly integrated with instruction is Robert's use of warm-up exercises to both check understanding and provide additional practice (described previously in the *Information* and *Use* sections). Here we describe an instance in which the integration of assessment and instruction is even tighter: the information gleaned is highly nuanced, and the action taken is highly responsive. In the following example, Janet uses warm-ups in a feedback loop in which specific information about students' difficulties is used to model a correct solution path, select appropriate future problems, and work closely with students who continue to show difficulty:

> I definitely go through [the warm-up task] again, go through the steps. "This is how you do this, this is what you do here, this is what you do here." And then in order for me to

> see that they understand it, I'll create another problem of my own very similar to that one. And then, if I do find one or two students that still don't understand, I will ask them what it was that they did not understand about this problem specifically. And that's the way I assess them with the daily warm-up exercises.

Here, the warm-up activity is clearly instructional, though Janet describes it as an assessment. She also seeks specific information about students' difficulties with the task, with the implication that she will respond to this information during the current activity.

Examples of the tight integration of assessment and instruction also came in teachers' discussion of peer assessment activities. In the following example, Nancy describes the rich instructional opportunities she observed in a peer assessment activity from her class, even though it was implemented in part to share the workload of correcting papers (here, the constructed-response Noyce tasks):

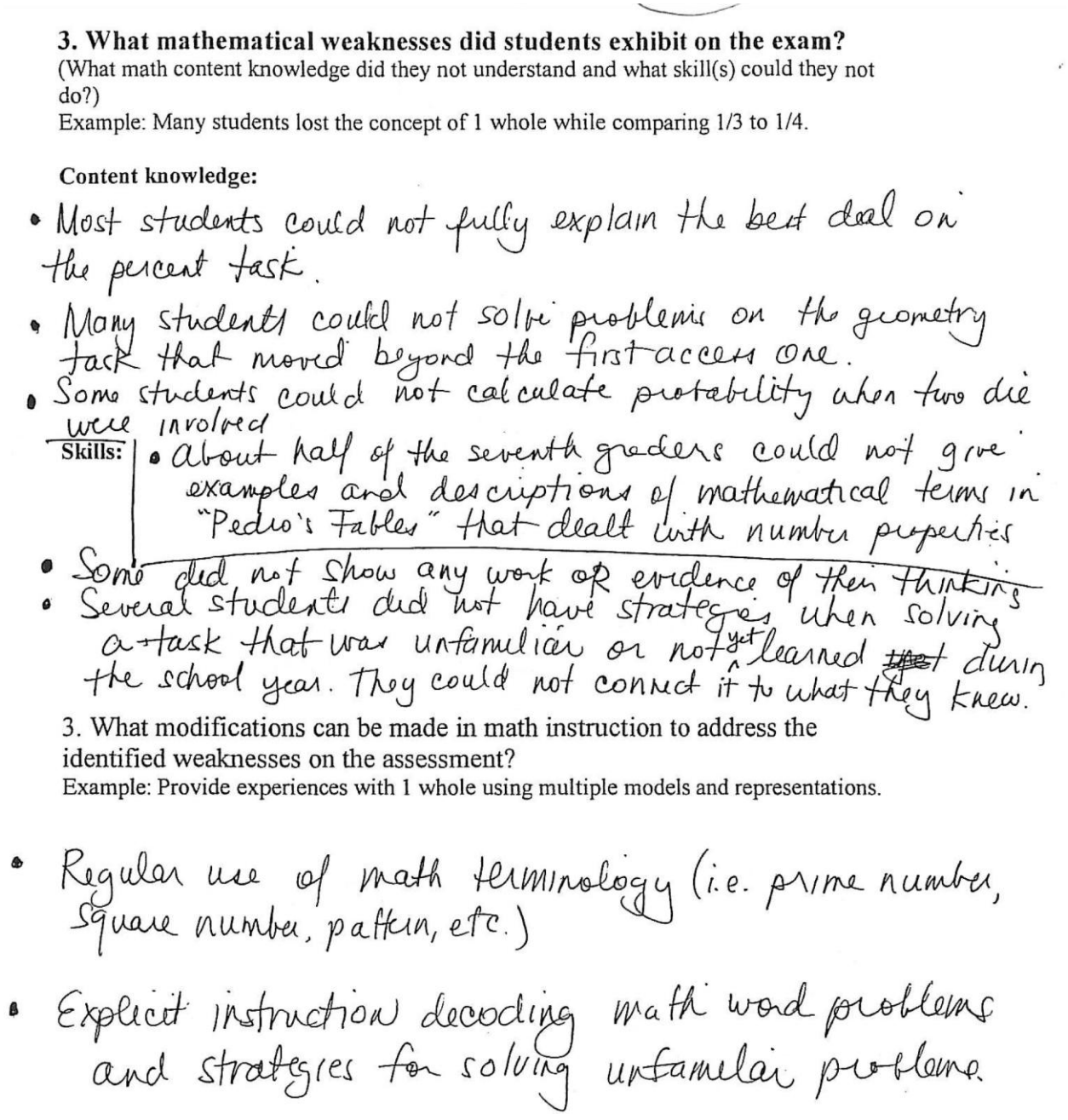
> Sometimes… I can't get to look at [the tasks], but I give them to everybody to—okay, I haven't been able to grade this formula yet, so let's talk about it and see where your differences are and explain to each other why you think it's different, and so maybe you can convince each other of who's right and who's wrong. And then you all come up with a general answer sheet. So that's a different way that I use it. So it's not really a quiz or a test or a grade or a number that I use, but from there I can see… you guys changed your answer on this. What was the discussion, what was the learning, what did you get out of this that you didn't know before? So they teach each other almost, which is great.

In a similar example, Leah reported using the review of a previous assessment as a forum in which the teacher and students together discuss a range of solution paths to assigned problems. Here, she reveals how this activity led to new knowledge about how to approach problems of a particular type:

> So they'll get in groups and do this together… and I'll try to make sure that there's at least one kid in the group who I know is a pretty abstract thinker and—and then I'll have them all grade it together. They're in groups and then we all share our grades and the explanations of why. And each group will be the leader, take turns on each different method and then talk about why they thought it was good or not. And it's always very fun, because you'll get brilliant ones that some group will give a really bad grade to. And it just brings up great discussions, and have these kids go, ―Oh! like they'd never even thought of it that way. A totally different method.

Opportunities to integrate assessment and instruction were not always teacher-generated. As part of the communal grading of the MARS assessments, teachers filled out a template in which they reported the score distribution for the group of students assessed, areas of strength and weakness demonstrated by the students, and suggested instructional

modifications to address the identified weaknesses. Figure 3 is an excerpt from a reporting template for a MARS assessment, as completed by Ann (the entire template is included in Appendix B):



3. What mathematical weaknesses did students exhibit on the exam?
(What math content knowledge did they not understand and what skill(s) could they not do?)
Example: Many students lost the concept of 1 whole while comparing 1/3 to 1/4.

Content knowledge:

- Most students could not fully explain the best deal on the percent task.
- Many students could not solve problems on the geometry task that moved beyond the first access one.
- Some students could not calculate probability when two die were involved

Skills:

- About half of the seventh graders could not give examples and descriptions of mathematical terms in "Pedro's Fables" that dealt with number properties
- Some did not show any work or evidence of their thinking
- Several students did not have strategies when solving a task that was unfamiliar or not yet learned ~~that~~ during the school year. They could not connect it to what they knew.

3. What modifications can be made in math instruction to address the identified weaknesses on the assessment?
Example: Provide experiences with 1 whole using multiple models and representations.

- Regular use of math terminology (i.e. prime number, square number, pattern, etc.)
- Explicit instruction decoding math word problems and strategies for solving unfamiliar problems

Figure 3. Excerpt from a Noyce reporting template.

This template works to integrate assessment in the traditional, summative sense (which describes students' current level of knowledge) with assessment in the formative sense

(which determines how to support students in achieving learning goals related to ongoing instruction).

Teachers for whom assessment and instruction were closely integrated reported selecting or designing assessment in order to learn very specific information about their students‘ understandings that would be useful for planning instruction. For example, Nathan discusses using clickers, in which students could ―click‖ (in real time) their response to a multiple choice question, as a way of gathering information about particular misconceptions:

> This is assessment is written to—that a lot of the incorrect choices are [represented]—and those are misconceptions. For example, a classic misconception in item number eight is that when you add mixed numbers, a lot of students don‘t get a common denominator and just add the whole number part, add the numerators and add the denominators. That‘s classic. So let‘s see—if you did that, it would be seven and four-fourteenths--which was choice C.

Jeff‘s use of linked items on weekly quizzes, described previously in the *Information* section, is another example of an assessment designed to elicit specific information for the purpose of instructional planning.

In the following excerpt, Jeff discusses the central role that the information he gathers from these quizzes plays in his classroom instruction:

> I plan my lessons based on [self-described formative quizzes] so I know going into the week how they did on their previous week's quiz. And then I know, oh, they've got this week, and move on, or I can go more in depth. Or there's just this one question they all got wrong, so I know I need to make sure I cover that concept better, but I can probably move on in other ways. It determines what I do the following week or two weeks or whatever.

In practices such as these, assessment is approached as a tool to refine and improve instruction, rather than simply an add-on to instruction. This thoughtful integration of assessment with instruction, using nuanced information in ways that are highly responsive to students‘ understandings, exemplifies our definition of ―true‖ formative assessment.

Research Question 3: Comparison of Findings Across the Three Formative Assessment Systems

To address our third research question, we looked for ways in which assessment practices varied among teachers using the Noyce, Formative-P, and Freudenthal systems. While we found examples of most of the categories of information gained from assessment and use of assessment in all three systems, there were some notable differences in the

prevalence of some of these findings from one system to another. While these patterns are interesting, it is important to note that we cannot make causal claims about these relationships (e.g., we cannot claim that a difference between the practices of teachers using two different assessments is definitely caused by the difference in the assessments). Our samples of teachers and events were too small, non-random, and unsystematic to draw strong inferences about the effects of particular formative assessment systems. We do, however, speculate about the manner in which some of the differences in the three assessment systems might be related to the variation in assessment information and use reported by the three groups of teachers.

Comparison of the information teachers obtained across the three systems. The reports of information obtained from various assessments were generally more similar between teachers using the Noyce and Freudenthal assessment systems than were those of either of these groups to the teachers using Formative-P. For example, more of the Noyce and Freudenthal teachers reported obtaining highly nuanced and detailed insights into individual students' understanding than Formative-P teachers. Three of the four Noyce teachers in our sample discussed examples in which they explored an individual student's thinking in great detail, obtaining nuanced insights; the fourth teacher discussed moderately nuanced information. Four of the six Freudenthal teachers gave examples of highly nuanced insights, one gave a moderately nuanced example, and one did not discuss individual students' work. Of the five Formative-P teachers, only one provided a highly nuanced example, two provided examples with moderate degrees of nuance, and two did not discuss individual students' thinking. There are many possible explanations for these differences; it is interesting to note how they might be related to the features of the three assessment systems. Consider, for example, the open-ended nature of the Noyce tasks and Ann's remark that these tasks allow her to ~~"s~~see" her students' thinking. The Freudenthal training aims to increase teachers' awareness of the variety of *types* of questions used for assessment, including those that call for analysis as opposed to simple recall. These features are central to the Noyce and Freudenthal assessment systems. The idea of exploring students' thinking was less prominent in the teachers' descriptions of the Formative-P system. This is somewhat surprising, because the Formative-P student assessments included elements that asked students to explain the reasoning behind key steps in solutions to problems. The teachers we interviewed who were using Formative-P seemed more engaged with the summative reports on their students' overall performance that they had received back; though they did report some analysis of student thinking associated with these reports. For example, one teacher explained how the analysis of groups of scores included discussions of why students might

have chosen a particular incorrect response, which would likely involve some exploration of students‘ thinking—especially in terms of common misconceptions.

Teachers in both the Formative-P and Noyce groups, but not the Freudenthal group, mentioned learning about their students‘ mastery of a variety of different topics from a single assessment; this reflects the designs of the assessment systems (both Formative-P and Noyce include a cumulative exam). This breadth of information from a single assessment, however, was more often reported by teachers using Formative-P than Noyce. Along with the Freudenthal teachers, the Noyce teachers more frequently discussed obtaining information relative to a narrow range of content that they had recently covered with their students. This also seems to fit the structure of the three programs: the Fruedenthal training is intended to address teachers‘ ongoing assessments connected to current content; the Noyce tasks are also often used along with related content currently being taught; and the Formative-P assessments are less frequent and cover basic principles that are related to a broad range of topics.

It makes sense that the design of these three assessment systems would have some effect on the information teachers reported they had obtained; given the nature of the assessments that were being used, while we cannot make causal claims, the patterns we observed are logical. This idea will be explored further in the *Discussion* section.

Comparison of teachers' use of information across the three systems. In talking to teachers about their assessment practices, we saw considerable variation in the responsiveness of the actions teachers took in using assessment information, as previously discussed in the *Use* section. Here, we consider the question: Were different assessment systems associated with different levels of responsiveness? We also compare these systems in terms of the previously discussed characteristics of ―true" formative assessment: purpose, cycle of use, and planned integration with instruction.

While we did find examples of the least responsive action to assessment information (i.e., to do no remediation in the face of evidence of student difficulty but move on to the next topic), such a small number of teachers reported this as a common practice that it cannot be related to any of the three systems. A somewhat more responsive approach to the use of information from assessments, and one which was more commonly reported among teachers in our sample, was to simplify the original lesson and re-teach it to the entire class, typically by simplifying examples and/or slowing down the pacing of the lessons. In this common pattern of action, which could be termed *simplify and repeat*, teachers did not identify or respond to specific elements of a lesson that seemed to be giving students the most trouble.

Rather, assessments seemed to prompt action only when a particular threshold number of students did not perform well. While teachers‘ practices across all three assessment systems included at least some evidence of this pattern of use, it was most common among the Formative-P group, with the majority of teachers (three of five) reporting this type of practice; two of the six Freudenthal teachers, and none of the Noyce teachers, mentioned practices that fit this characterization. Another moderately responsive action was to review the completed assessment, discussing the responses with the class. Though teachers in all three groups reported doing so, the Formative-P teachers typically described modeling correct solutions for their students, while the Noyce and Freudenthal teachers described a discussion-based format.

A less common (though more responsive) approach to the use of assessment information was to use information from assessments to target *particular areas of student difficulty* that were common among students within the class (e.g., revising the original lesson to target student difficulties in that area). While this type of response to information from assessment was targeted in terms of content focus or learning proclivities, it generally did not involve the planned selection or design of assessments in terms of what they might reveal about student learning. Rather, teachers took advantage of informational opportunities that the assessments happened to provide. This pattern of practice was most common among the Freudenthal teachers, being reported by four of the six. Two of the four teachers using Noyce, and one of the five using Formative-P, reported this practice, which could be termed *revise and reteach*. Another way teachers targeted particular areas of difficulty was to require students to submit corrections to their completed assessments. Though teachers in all three groups mentioned this practice, it was described somewhat differently by Noyce and Freudenthal teachers (who typically required students to explain their responses) than by Formative-P teachers (who required corrections but not explanations).

The ideal of ―true" formative assessment represents the most responsive use of information about student learning. As previously mentioned, it involves a close feedback loop between instruction and assessment, such that information from assessment is regularly used to inform planning and on-the-fly adjustment to instruction. Moreover, the assessment opportunities themselves are *planned for* in terms of their affordances to elicit insight into student thinking. Though we found relatively scant evidence of this pattern of practice in our study, the teachers who did describe assessing in this manner were all in the Freudenthal and Noyce districts, with half of the teachers in each of these groups reporting practices that met these criteria.

As with the information teachers gleaned from assessments, the ways in which they used the information was most similar between Noyce and Freudenthal teachers. However, it is more difficult to connect the ways teachers choose to use the information to features of the assessment systems themselves. Teachers directly gain information from the assessments; thus, the information is connected to the features of the assessments to some degree. It is less clear what the connections might be between the assessment features and information use, which is one more step removed from the assessment itself (administer assessment --> obtain information --> use information). They may be connected indirectly by the types of information the various assessments provide, which might afford or constrain the ways in which teachers respond to this information.

Discussion

Formative assessment has been heralded as a potent tool for improving instructional quality and student learning. Districts that are interested in promoting formative assessment face a variety of options ranging from systems with pre-designed tests and assessment events to training that focuses on developing educators‘ own formative assessment expertise. The purpose of this study was to increase our understanding of the ways teachers interact with various formative assessment systems. In particular, we sought to describe more completely the kinds of information teachers glean from formative assessment, how they adapt their behaviors in response to this information, and how these factors are influenced by the characteristics of the approaches to formative assessment they are using.

Limitations

The study was designed to portray the range of teacher interactions with formative assessment systems and to provide rich exemplars of different kinds of practices. Neither the sample of formative assessment systems nor the sample of assessment events we explored is representative of the full universe. Though we provide some information about prevalence within our sample, we fully acknowledge that these cannot be generalized to all teachers, all districts, or all formative assessment systems. Nevertheless, we think these exemplars provide valuable information about the delicate interface between assessment systems and teachers‘ applications in their classrooms. It is important to understand how the theoretical advantages of formative assessment play out against the practical constraints of classrooms.

Variation in Assessment Information and Its Use

We found that teachers draw different kinds of inferences from the evidence provided by formative assessment, sometimes making very coarse judgments about student understanding and sometimes more refined judgments. Most commonly (in this sample),

teachers seemed content to know whether the majority of students ―got it" or not. For whatever reason, in these instances, it was sufficient for their purposes to make a binary decision (i.e., did their students possess the target knowledge and skill or not)? Less often (in this sample), teachers used the assessment results to make subtle distinctions about student learning of the targeted concept (e.g., whether students had partial knowledge; which sub-skills or sub-processes were understood and which were causing problems; which students mastered the lesson and which did not, etc). We distinguished three levels of nuance, but believe these categories are less important than the general principle that teachers choose how much information to extract from assessment results—sometimes being satisfied with gross generalizations and at times being interested in more particular insights. A formative assessment system may have the potential to inform highly nuanced insights into student understanding; yet, for a variety of reasons we do not fully understand, teachers may not draw out those insights.

Though this process is not entirely understood, we presume that the differences in the active interpretation of assessment results arise from many factors. From the remarks and examples in this study, we have identified three contributors to this variation: 1) teachers' knowledge and skill with respect to both assessment and mathematics; 2) the specific purpose(s) of the assessment; and 3) the instructional context in which teachers operate. As noted, there was variation in nuance within teachers as well as between teachers. Within-teacher differences (i.e., drawing binary conclusions in one instance and highly nuanced conclusions in another) can be partly explained by subtle differences in purpose between the assessments. If a teacher believes students have mastered the current topic and are ready to move on, she may be simply seeking confirmation of this at a yes-or-no level. If, on the other hand, the class is still immersed in the topic, she may be interested in learning about specific areas of student strength or difficulty that she can address in her instruction. The idea of purpose is also connected to teachers' knowledge of formative assessment (this will be discussed in the *Professional Development* section). Within-teacher differences might also be explained in some instances by difference in classroom context, including variations in student mastery of specific content, pressures exerted by pacing guides and external events, etc. If the teacher is running out of time during the period set aside to cover the distributive property, then he or she may be limited to making a broad judgment about overall mastery. It is also possible that within-teacher differences could result in some cases from a teacher's own variation in expertise between math topics or assessment types. A teacher with deeper knowledge in the area of rational numbers than geometry, for example, would be better equipped to examine students' own understanding of the former topic than the latter; one

who has just been introduced to the idea of using a warm-up assessment activity might not glean as much information from such an assessment as from a weekly quiz he has been using for years.

Between-teacher differences (i.e., one teacher describing primarily binary distinctions while another describes more nuanced distinctions) might also be plausibly explained by any of the three reasons just mentioned. Contextual factors such as curriculum pacing and class size can vary considerably among districts, schools, and even classrooms (e.g., in the case of class size). Several teachers in our study noted limitations pertaining to the information they were able to obtain from assessments that resulted from these factors. There are also differences in the way teachers view the purposes of assessment. Some teachers are more oriented to summative and grading purposes; others are more oriented to instructional purposes. These distinctions, in turn, can be connected to differences in school/district leadership and philosophy and to teacher knowledge. Differences in teachers‘ understanding of assessment and mathematics would clearly affect the differences in information they gain from math assessments. For instance, a teacher with deep mathematical knowledge will find the results of the quiz about equations full of interesting insights into students‘ understanding of rational number equivalence and inverse operations. On the other hand, a teacher with shallower knowledge will see these same results as only a measure of their ability to follow procedural steps. Likewise, a teacher with a strong understanding of the various purposes of assessment and assessment techniques will likely draw more information from assessment than one with less expertise in this topic.

Our data do not allow us disentangle these factors further. However, we think it is important to try to understand why teachers‘ assessment information function varies to such a degree. Is this an active choice that demonstrates strategic behavior? If so, then sub-optimum information gathering might be enhanced through training. Alternately, are teachers overwhelmed by events that constrain their ability to extract useful information from assessment? If so, a different approach might be needed to improve the situation.

We also found interesting variation in teachers‘ use of assessment information, which ranged from deciding to move on to the next topic with no further instructional activities (even in the face of identified deficiencies in learning) to engaging in highly-focused, highly-responsive learning activities designed to enhance student understanding of specific concepts and procedures. It was quite common for teachers to tell us that after a formative assessment (quiz, check sheet, etc.) they had decided to do nothing more than originally planned and move on to the next topic. It was also common for teachers to describe instances in which deficiencies identified through assessments prompted them to go back and cover the material

in a similar way as before. It was less common for teachers to describe situations in which assessment results prompted them to re-teach material in a different way or to focus an additional lesson on an aspect of a skill or procedure that had been revealed as difficult for students. As in the case with assessment information, what is important here is not the classification of these events into categories; rather, the significant point is the general principle that teachers were selective in their response to insights from formative assessment. Thus, at times, teachers decided that the results did not warrant a change in plans. At other times, instructors opted to spend more time on instruction—sometimes in new and more focused activities. Formative assessment was the trigger for these instructional choices, but we do not fully understand the relationship between the assessment results and teachers' subsequent actions.

We speculate that teachers' actions were based on a combination of their instructional skills and knowledge as well as their instructional context. A teacher with a limited repertoire of instructional options for a given topic is less likely to offer a new learning activity than one with a broader skills base. A teacher required to follow a rigid pacing plan or an educator that feels pressure to cover material, which will be included in a high-stakes test, is less likely to take the time to re-teach material than one who has greater control over classroom time or less accountability pressure. It is important to gain a better understanding as to why teachers use the results of formative assessment in different ways. If teachers are constrained by their own knowledge it suggests different improvement actions than if they are constrained by external factors. These insights are also relevant to the design and adoption of formative assessment systems. Systems with fixed tasks and schedules would be more useful in some contexts and with certain teachers; whereas, systems that offer less structure and greater flexibility (with respect to the choice of assessment tasks and the schedule of administration) would be a better fit for other teachers.

Lastly, there was a relationship between the information variable and the use variable. For example, more-nuanced insights were associated with more-refined instructional uses (as discussed previously in our description of ~~"t~~rue" formative assessment). This makes sense because more nuanced insights into students' understanding are a necessary condition for instructional actions that are more responsive to their understanding. Binary judgments most frequently formed a basis for less-responsive actions, such as re-teaching a lesson in a similar manner as it had been taught the first time. However, binary judgments were sometimes used in a planned and moderately responsive manner (such as in the case of a unit pre-test to determine which sub-topics a teacher needed to address and in what depth).

Professional Development and Formative Assessment

We suggested that teachers' differing levels of mathematics and assessment knowledge might explain some of the variations in information and use. To explore this possibility, we analyzed the teachers' responses to the interview questions we had asked them about professional development related to the formative assessment systems they were using. These questions dealt with the content and emphasis of the professional development; the teachers' responses revealed both similarities and differences among the three programs. Though we can no more make causal claims about the effects of these professional development programs than about the effects of teachers' use of a particular assessment system, it is interesting to consider the somewhat different emphases in these professional development efforts, as described by the teachers, in light of the variation in assessment information and use we found among the three groups of teachers.

In the Freudenthal district, participation in professional development was optional at both the school and teacher levels.[5] Two of the six teachers we interviewed in this district did not mention the Freudenthal training; however, they did discuss other professional development programs related to assessment that were also offered in the district. Those who discussed Freudenthal described after-school sessions in which they presented assessments they had used in their classes. Some teachers in the Freudenthal district also shared that they had received feedback from session leaders as well as other teachers; others discussed the types of questions they were using with respect to the Freudenthal assessment pyramid. The Freudenthal teachers reported that the professional development also emphasized informal assessment strategies, especially formative assessment. None of the teachers in the Freudenthal groups mentioned any work on instructional strategies other than questioning; the professional development appears to have been solely focused on assessment.

All of the Formative-P teachers reported participating in professional development, which was a required part of the experimental study described in the introductory section. These teachers described an initial session introducing the program, presentations on the three units of instruction and use of the instructional materials, and a session after each of the four assessments were administered (in which they analyzed the results for groups of students). Judging by the teachers' remarks, their experiences with the Formative-P professional development appear to have been almost equally focused on instruction and assessment. In general they spoke highly of the professional development, reporting that they

[5] In the case of Freudenthal, the ~~"p~~rofessional development program" is essentially the same as the ~~"a~~ssessment system;" thus we note some redundancy between this section and the *Comparisons* section above.

learned to present information in ways more easily understood by students. They spoke somewhat less about the analysis of assessment results than they did about the sessions on instruction. In addition to examining a breakdown of students' scores by item, one teacher mentioned discussing ways to better teach students based on these results, and one mentioned discussing reasons for specific incorrect responses.

All four of the teachers using the Noyce system reported participating in professional development related to that program. They described monthly full-day sessions in which they practiced solving problems similar to the Noyce tasks and discussed possible student responses, as well as communal scoring of the quarterly assessments they gave their students. Teachers noted that the monthly sessions addressed both content knowledge and pedagogy; furthermore, they mentioned that the scoring sessions included analysis of student responses and produced "toolkits" of common student errors. In addition, two teachers mentioned participating in week-long summer institutes on problem solving, questioning, and the use of manipulatives. Thus, while this program appears to address instruction to some degree, the majority of the work is apparently centered on Noyce-like, open-ended assessment tasks.

The Noyce and Freudenthal programs were structured quite differently; however, the fact that their professional development efforts were both centered around in-depth analysis of assessment tasks—teachers' own assessment questions in the cases of Freudenthal, and externally-developed open-ended tasks in the case of Noyce—is related to the similarities in the teachers' reports of information they gained from their assessments and the ways they used that information. It seems reasonable that an intensive focus on assessment tasks and questions (including elements such as communal scoring of open-ended tasks or a focus on the use of questions requiring "analysis" responses) could result in an increased tendency to seek out nuanced information from assessments.

Conclusions

We were somewhat disappointed that we did not find more instances of highly nuanced insights from assessment, highly responsive uses of information, and patterns of practice that we could characterize as "true" formative assessment. Though we did find examples of each of these characteristics, they were the exception rather than the rule among the data we obtained from our teacher interviews. Upon further reflection, however, we also found cause for optimism. Formative assessment is widely recognized as a powerful integration of assessment with instruction that has great potential to improve student learning. Having been in the spotlight for a little more than a decade, it is a topic that is relatively new to education research and practice; thus, it is a field that is still developing rapidly. In education, changes

in mindset and practice take time, and formative assessment represents a dramatic change in both mindset and practice (as compared with the traditional view of assessment as a summative tool used mainly for the purposes of grading). When regarded in this light, our findings reveal significant progress toward a more formative view of assessment. All of the teachers we interviewed reported regularly using the results of classroom assessments, such as warm-ups and quizzes, to adjust their instruction. Most educators described informal types of assessment they used to monitor student learning during instruction, as well as near-term actions they took with individual students or the whole class on the basis of these assessments. Some of the teachers reported periodic, brief assessments given for the sole purpose of planning their following instructional steps rather than grading. Most instances of these practices fall short of our definition of ~~"t~~rue" formative assessment, and many involved binary rather than highly nuanced information; however, they all contain formative elements and represent a departure from traditional, summative assessment practices. While formative assessment may not yet be the norm, it appears to be gaining a foothold.

We focused our study on three types of formative assessment to allow us to see how features of the assessments were associated with the information they provided and the manner in which this information was used. Our sample was too small to make any strong comparative judgments about these three systems. Nonetheless, three general observations seem warranted. First, for each assessment system there was considerable variation among teachers in the information they obtained and how they used it. Adopting Freudenthal, Noyce, or Formative-P did not eliminate variation in the insights teachers obtained from results or how they used the information. Lacking information on teacher behaviors prior to the adoption, we cannot say whether any of these systems reduced between-teacher variation and made practices more similar. Second, at the time we talked with Formative-P teachers the system was being implemented for the first time, whereas the Noyce and Fredenthal systems had been in place for a few years. Greater familiarity with the formative assessment system did seem to be accompanied by more integrated use during the school year. Only Formative-P teachers tended to compartmentalize the assessment system to particular time periods and topics, even though they were free to use the instructional materials as they saw fit (aside from the prescribed assessment timeframes). In many ways, at the time we observed, Formative-P had many of the characteristics of an interim assessment system (see companion paper, Shepard et al., in press). Finally, teachers seemed to find it easier to incorporate the systems that had pre-existing assessments (Noyce and Formative-P) than the system that put the burden for assessment design on their shoulders. This should not be surprising; in fact, it may explain the popularity of packaged formative and interim

assessment systems over the kinds of professional development that empower teachers to incorporate assessment as an integral feature of instruction. Again, a clearer understanding of these patterns would be useful for designing formative assessments, choosing which to implement, and training teachers to use them more effectively.

While not definitive by any account, these results are potentially useful for a variety of purposes. The results from this study can aide teachers, administrators and other education stakeholders in deciding which formative assessment systems to adopt, planning for the implementation of formative assessment and providing adequate training for teachers, designing formative assessment systems that better meet teachers‘ needs, setting realistic expectations for the impact of formative assessment systems on a large scale, and lastly, understanding the impact of formative assessment in a particular context.

References

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-148. Retrieved from http://www.pdkintl.org/kappan/kbla9810.htm

Burch, P. (2006). The new educational privatization: Educational contracting in the era of high stakes accountability. *Teachers College Record, 88*(2), 129-135.

Davidson, K., & Frohbieter, G. (in press). *District adoption and implementation of interim and formative assessments*. (CSE Technical Report). Los Angeles, CA: University of California, National Center for Researcher on Evaluation, Standards, and Student Testing (CRESST).

Foster, D. & Noyce, P. (2004). *The Mathematics Assessment Collaborative: Performance testing to improve instruction*. Palo Alto, CA: Noyce Foundation.

Foster, D., Noyce, P., & Spiegel, S. (2007). When assessment guides instruction: Silicon Valley‘s Mathematics Assessment Collaborative in assessing mathematical proficiency. In A. Schoenfeld (Ed.), *Assessing Mathematical Proficiency* (pp. 137-154). Cambridge, MA: Cambridge University Press.

Greeno, J .G., Pearson, P. D., & Schoenfeld, A. H. (1996). *Implications for NAEP of research on learning and cognition*. Report of a study commissioned by the National Academy of Education, Panel on the NAEP Trial State Assessment, Institute for Research and Learning. Stanford, CA: National Academy of Education.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and synergies. *Curriculum Journal*, *16*(2), 207-223.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library*. London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W. & James, M. J. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education*, *4*(3), 365-380.

Linn, R. L. & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. N. Baron and D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth yearbook of the National Society for the Study of Education, Part 1) (pp. 84-103). Chicago, IL: National Society for the Study of Education (distributed by the University of Chicago Press).

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making*. Santa Monica, CA: RAND Corporation, OP-170.

Massell, D. (2001). The theory and practice of using data to build capacity: State and local strategies and their effects. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states*. Chicago, IL: University of Chicago Press.

National Council of Teachers of Mathematics (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

Phelan, J., Choi, K., Vendlinski, T., Baker, E. L., & Herman, J. L. (2009). *The effects of PowerSource© intervention on student understanding of basic mathematical principles*. (CRESST Report 763). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston, MA: Kluwer.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation*. Chicago, IL: Rand McNally, American Educational Research Association monograph series on evaluation, No. 1.

Shepard, L., Davidson, K., & Bowman, R. (in press). *How middle school mathematics teachers use interim and benchmark assessment data*. (CSE Technical Report). Los Angeles, CA: University of California, National Center for Researcher on Evaluation, Standards, and Student Testing (CRESST).

Shepard, L. A., Hammerness, K., Darling-Hammond, L., Rust, F., et al. (2005). Assessment. In L. Darling-Hammond and J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 275-326). San Francisco, CA: Jossey-Bass.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Stein, M. & Bassett, E. (2004a). *Staying ahead of the curve: A value chain analysis of the K-12 assessment market*. Boston, MA: Eduventures, Inc.

Stein, M. & Bassett, E. (2004b). *Uncovering K-12 professional development opportunities*. Boston, MA: Eduventures, Inc.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy and Practice*, *11*(1), 49-65.

Appendix A:
The CRESST Project on Interim and Formative Assessment

Purpose and Sample

This study was part of a broader assessment project undertaken from 2005 to 2010 by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The centerpiece of that project was a randomized controlled trial of a formative assessment system developed by CRESST called PowerSource.© To augment the CRESST experimental study, researchers from the University of Colorado, the RAND Corporation, and Stanford University conducted a qualitative study of middle school mathematics teachers' use of a variety of formative and interim assessments (including Formative-P, which was an early implementation of PowerSource©). The purpose of the qualitative study is to explore middle school math teachers' use of data from formative and interim assessments as well as the relationship between assessment types and instructional uses.

Methods

Distinguishing among types of assessments. When considering distinctions among types of assessments that might be considered formative, we began with the definitions of interim and formative assessments posed by Perie, Marion, and Gong (2007, p. 5). These authors state that interim assessments are those ~~"a~~dministered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level," whereas ~~"f~~*ormative assessment* is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes." The authors note also that results are aggregated across classrooms, schools, and/or districts for interim but not formative assessment. Reviewing the literature, we observed a continuum of assessment types that fit these two definitions (according to how closely they are connected to classroom instruction) (see Figure A1). At the distal end of the continuum are interim assessments with no formative use, as the content being tested is not typically the same as that currently being taught. At the proximal end are formative assessments that cannot be aggregated to serve the purposes of interim assessment, as they are not standardized. This paper focuses on the more proximal piece of this spectrum, while a companion paper (Shepard, et al, in press) addresses the distal range.

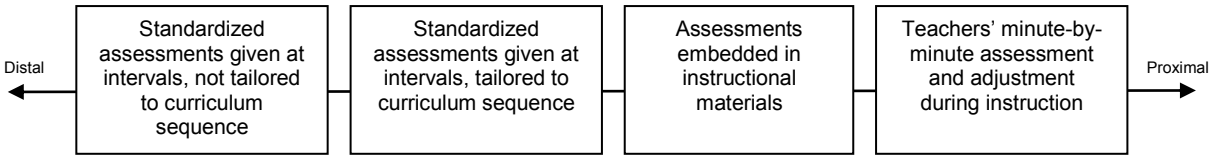


Figure A1. A continuum of interim and formative assessments.

Sampling. Our sampling process was purposive and also one of convenience (Maxwell, 2005). We sought to include assessments at all locations on the continuum that were in prevalent use in the three geographic regions in which the research teams were located, and where possible, to select assessments that were in prevalent use nationally. We chose the following assessments for the four categories of assessments shown in Figure A1:

1. Interim assessment system not linked to curriculum: Northwest Evaluation Association Measure of Academic Progress System (NWEA/MAPS)
2. Interim assessment system tailored to curriculum:
    a. District-developed interim (or benchmark) assessments in three districts
    b. Evans-Newton developed interim assessment, or District Wide Assessment (DWA)
3. Assessment system embedded in instructional materials:
    a. Formative-P
    b. Noyce
4. Classroom formative assessment: Teachers‘ classroom assessment associated with a district professional development program aimed at formative assessment.

Selection of Districts and Respondents. We began by identifying districts that were using the assessments we had selected. In addition, we contacted the appropriate administrator to obtain permission for the research. We also asked the administrator to identify two schools that were using the assessment of interest; then, either a member of our team or the administrator contacted the principals of these schools to request their participation. Principals who agreed to participate were asked to help us recruit three teachers to be interviewed for the study. We were not always successful in recruiting the number of schools and teachers sought for each district. Our final sample of eight assessment systems, ten districts, and 42 teachers is shown in Table A1.

Table A1.

Study Sample

District	Number of teachers	Assessment type(s)
Washington	6	Northwest Evaluation Association Measure of Academic Progress System (NWEA/MAPS)
Madison	6	District-developed interim/benchmark assessment
Taylor	6	NWEA/MAPS
Jackson	6	Professional Development on Formative Assessment (Freudenthal)
Burlington	4	Evans-Newton developed Interim assessment, called District-Wide Assessment (DWA)
Sinclair	3	Formative-P and district-developed interim assessment
Pittsfield	4	District-developed interim assessment
Adlington	3	Formative-P
Prairie	3	Noyce
Alta	1	Noyce

The data drawn from these districts were used as the basis for three related sets of analyses, one focusing on district intent (Davidson and Frohbeiter, in press), one on information and use of interim assessments (Shepard et al., in press), and the current study on information and use of formative assessments.

Appendix B:
Reporting Template—Communal Grading of Noyce Assessments

ıentary School District
Math Assessment
ınalysis of Student Work

School

Grade Level 7 , 2007-2008

1. Data Analysis:

Rubric Score	30-40	20-29	10-19	0-9
Number of Students Receiving a Particular Rubric Score / Number of Students Taking the Exam	15 / 190	78 / 190	82 / 190	15 / 190
Percentage	8%	41%	43%	8%

2. What mathematical strengths did students exhibit on the exam?

(What math content knowledge did they understand and what skill(s) could they do?)
Example: Most students were able to determine that 1/3 is greater than 1/4.
Content knowledge:

3. What mathematical weaknesses did students exhibit on the exam?
(What math content knowledge did they not understand and what skill(s) could they not do?)
Example: Many students lost the concept of 1 whole while comparing 1/3 to 1/4.

Content knowledge:

- Most students could not fully explain the best deal on the percent task.
- Many students could not solve problems on the geometry task that moved beyond the first access one.
- Some students could not calculate probability when two die were involved

Skills:

- About half of the seventh graders could not give examples and descriptions of mathematical terms in "Pedro's Fables" that dealt with number properties
- Some did not show any work or evidence of their thinking
- Several students did not have strategies when solving a task that was unfamiliar or not yet learned ~~that~~ during the school year. They could not connect it to what they knew.

3. What modifications can be made in math instruction to address the identified weaknesses on the assessment?
Example: Provide experiences with 1 whole using multiple models and representations.

- Regular use of math terminology (i.e. prime number, square number, pattern, etc.)
- Explicit instruction decoding math word problems and strategies for solving unfamiliar problems.
- Solving computational math problems that focus on skill IN A REAL-WORLD CONTEXT.
- More opportunities for solving percent problems, probability, and areas of complex shapes in all