"They come to the test, and there is nothing to fold": Teacher views of large-scale assessments and classroom context

Caroline E. Parker, Education Development Center, Inc.
Susan Saxon, Educational Alliance at Brown University

January 2007

Support for this investigation was funded in part by Grant CFDA #84.368 of the U.S. Department of Education, Office of Elementary and Secondary Education, awarded to the Rhode Island Department of Education on behalf of the New England Compact. The opinions expressed in this paper do not necessarily reflect the position or policy of the Department of Education, and no official endorsement should be inferred.


Table of contents

Introduction

The last 10 years have seen rapid changes in the development of large-scale assessments. The No Child Left Behind Act of 2001 (NCLB) requirements that districts and schools measure student achievement, including that of students with disabilities and English language learners (ELLs), have created pressure to improve the ability of large-scale assessments to accurately measure student achievement and to ensure access for all students. To this end, assessments using universal design principles promote increased access through various means, such as font size, use of blank space, and use of illustrations in test construction (Thompson, Johnstone, Anderson, & Miller, 2005; Universally Designed Assessments, "Universally Designed Assessments," 2005). States have also increasingly standardized their use of accommodations to increase access (Bolt & Thurlow, 2004; Edgemon, Jablonski, & Lloyd, 2006; Ysseldyke, Thurlow, Bielinski, House, Moody, & Haigh, 2001). Other research has led to the development of new English language proficiency tests focusing on academic English acquisition in four domains (listening, speaking, reading, and writing).

Yet, even with all of these changes and the resulting increase in access for many students, educators argue that the current assessment structure as mandated by NCLB still does not reliably measure the achievement of *all* students (Almond, Quenemoen, Olsen, & Thurlow, 2000; Quenemoen, Thompson, Thurlow, & Olsen, 1999; Wiener, 2006). A significant number of students may not be well-served by the current generation of large-scale assessments inasmuch as the assessments do not reliably measure the achievement of these students. These gaps in the assessment structure are not well understood, and there is little research on either the gaps or the students who fall into them.

This paper forms part of a larger study of students in the assessment gaps. The study looked at the issue of the assessment gaps from multiple positions and through multiple methods. Teachers were interviewed to better understand their views of large-scale assessment and students in the gaps; large-scale assessment data were analyzed to identify students in assessment gaps; student questionnaires, completed as part of the large-scale assessment, were analyzed to understand some aspects of the student perspective; and multiple assessment modalities were piloted and analyzed. Each paper addresses one of five overarching research questions:

1. Who are the students in the gaps?
2. Of all students who are not proficient, how can states identify those who are in the gaps?
3. What are the attributes of students in the gaps, and how do these students perform?
4. What issues in the assessments themselves contribute to the gaps?
5. Are there specific aspects of the multiple-choice items (complexity, presentation) used in state assessments that contribute to the assessment gaps?

The study described in this paper focuses on the fourth question, looking at the issues in large-scale assessments that contribute to the gaps. This exploratory study looks at the question from the point of view of middle school mathematics and special education teachers. We chose to begin with teachers because they deal with students daily in the classroom, and they are the ones most likely to identify any discrepancies between the day-to-day classroom-demonstrated proficiency of their students and the large-scale assessment results. A second research paper using the same data addresses the first question of the larger study: *Who are the students in the gaps?*

This study takes place in the context of the purpose of large-scale assessment within NCLB: accountability for schools. If there are students who are working proficiently at grade level, but this is not being captured by the assessment, the schools will be unfairly judged as not meeting adequate yearly progress. If there are students who are working well below grade level but who are progressing in their learning, a grade-level assessment will not show that progress, and again, the schools will be unfairly judged. Thus, the stakes are high for getting assessments right.

Our exploratory study sought to more carefully identify those factors that contribute to students not being served well by the large-scale assessments, including but not limited to students with disabilities and ELLs. We sought to describe teachers' understandings of their own students' achievement in eighth grade mathematics, both in the classroom and on the large-scale assessment.

Methods

Twenty-three mathematics teachers, 14 special education teachers, and 3 administrators with special education or mathematics expertise from the four New England states were interviewed, for a total of 40 participants. The majority were eighth grade teachers, except in New Hampshire, where the middle-level 2004 mathematics assessment was administered in sixth grade. Interviews were conducted using an open-ended protocol, and access to specific student scores was provided, when available, as a starting point for discussion.

The interviews for this study were conducted in the spring and early fall of 2005, before the implementation of the NCLB requirement that students be tested annually in grades 3–8. In three of the states, the large-scale assessment has been totally restructured, and thus it should be noted that the interviews refer to the previous generation of assessments.

As an exploratory study, and because of time constraints, we decided that it would be best to choose our sample based on convenience. State education officials provided a list of three or four schools in each state whose teachers and administrators might be willing to participate in the study, and that reflected the diversity of schools in their state: large, medium, small, urban, rural. We then contacted the schools directly and set up the interviews. In some cases, the teachers had access to student assessment results, though in others they did not. In each school, all eligible teachers (mathematics or special education teachers at the appropriate grade level) were invited to be interviewed (see Appendix A for the interview protocol). In the smaller schools, all eligible teachers participated; in the larger schools, a selection of teachers participated. The following table provides an overview of the participating schools.

Table 1. Participating schools

	Urban status	School size	Grade span	% Caucasian	% free/ reduced -price lunch	Other	Teachers		
							Math	SPED	Other
State 1, school 1	Suburban	933	6–8	88	39		3	1	
State 1, school 2	Suburban	630	6–8	93	17		2	1	
State 2, school 1	Rural	579	7–8	98	35		2	2	
State 2, school 2	Rural	388	6–8	97	20		2	2	1 admin
State 2, school 3	Rural	291	7–8	99	23		3	2	
State 3, school 1	Rural	370	7–8	97	17	8% SPED	4	2	
State 3, school 2	Rural	427	K–8	96	8		1	1	
State 3, school 3	Urban district	~550	6–8				3	1	1 science, 1 admin
State 4, school 1	Small town	600	6–8	94	39	3% ELLs	3	2	

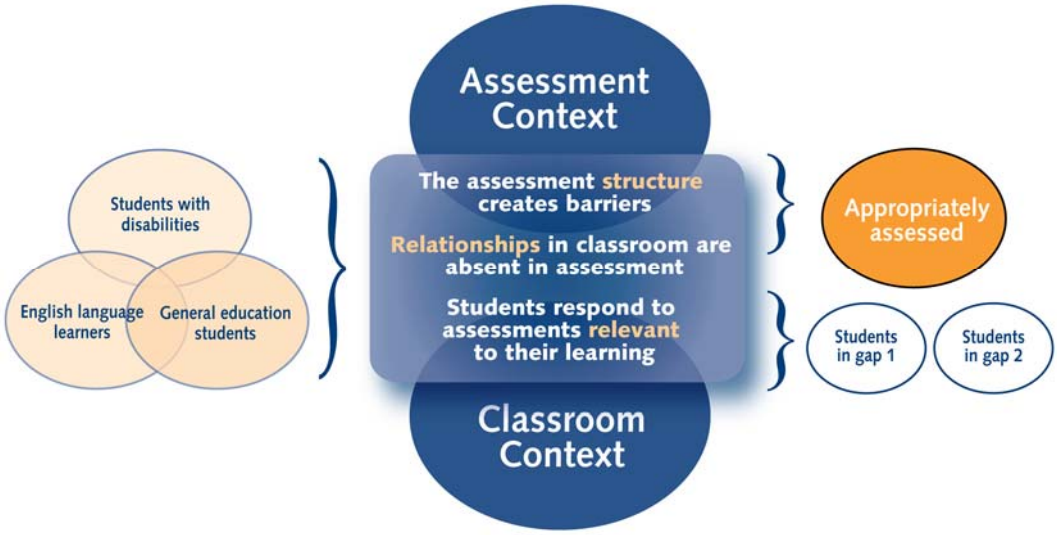
Interviews were conducted by four people, based on an open-ended interview protocol. They were then transcribed and analyzed, using qualitative software. Initial codes were developed based on both the literature and extensive meetings held with the project management team. These initial codes included "opportunity to learn," "learning style," "test stress," and "motivation," among others. Subsequent recoding led to new categories, which were then checked through recoding of interviews. Coding was done by the two senior researchers, who regularly shared coding strategies, coded each other's interviews, and redid coding based on emerging categories. As a result of this coding process, two major coding categories were developed: "student characteristics" and "assessment characteristics." These two categories led to the development of two research papers.

Because we chose to interview only teachers for this study, we recognize that there are limitations in the conclusions we can draw. We specifically sought to understand the teachers' viewpoint, but by limiting ourselves to this viewpoint, we do not have any data that would indicate teacher bias, such as student interviews classroom observations, or a systematic analysis of student assessment data.

Findings

Based on the results of the interviews in this study, and together with the other parts of the study, we have developed a theoretical framework, demonstrated in Figure 1. On the far left side are three groups of students: students with disabilities, ELLs, and general education students. When these three students are faced with the large-scale assessment, some of them are negatively affected by differences between the classroom and the assessment context. These differences are the focus of this paper and include the structure of the assessment, the differences between teacher practices and the assessment, and the lack of relevance of the assessment to student learning. Depending on the particular attributes of each student, these challenges result in them being affected by either the first or the second gap in the assessment system, as demonstrated at the far right of the framework.

Figure 1. Explaining the gap between classroom achievement and assessment results



This paper looks at the difference between the classroom context and the assessment context, and the gaps that this difference creates, from the teachers' point of view. Students in both assessment gaps are affected by structure, relationships, and relevance, though in different ways. Students in the first gap perform proficiently in the classroom but are stymied by the different context of the large-scale assessment. Students in gap 2, already performing below grade level, need scaffolding to show what they do know; these students in particular suffer when the assessment context has no relevance to what they learn in class.

Differences between assessment and classroom context

When we decided to talk to teachers about the large-scale assessment, we weren't sure what to expect. We thought that teachers might talk about tests not being good for all students, or that they might argue that the assessment isn't a fair measure of what students know and can do. We did hear both of these comments, but we also heard more: We heard that the structure of the assessment is problematic, we heard that person-to-person relationships are critical in the assessment process, and we heard that students need to see the relevance of an assessment in order to demonstrate what they know. We also heard, interestingly, that in many cases, the mathematics assessment is an accurate measure, from the teachers' view, of what their students know and can do. Most students who do not demonstrate proficiency on the assessment also do not demonstrate proficiency in class, or have not been exposed to the material on the assessment.

The assessment experience differs significantly from the classroom context in three ways:

- The structure of large-scale assessment creates barriers to demonstrating proficiency
- Teacher practices promote student demonstration of learning in ways that large-scale assessments do not
- Students respond better to assessments that are relevant to their learning

We found that different student characteristics interact with the classroom/assessment contextual differences and that none of these characteristics are mutually exclusive. The relationship between various factors, including student characteristics and classroom context, together create a context for student learning.

Structure: Assessment design creates barriers to demonstrating proficiency

NCLB requirements explicitly state that the development of large-scale assessments should include good practices of universal design for learning. However, there are at least two ways to think of universal design. One is to develop a single assessment that includes good practices to allow access to as many students as possible, through such techniques as font size and spacing. The second is to allow students to demonstrate their learning in the way that works best for them, which may not be through a sit-down assessment. Having access to multiple modalities of an assessment may provide an opportunity for more students to show what they know and can do. The current large-scale assessment systems in the four states in this study do not provide alternatives for administration,[1] and teachers described their frustration with this limitation:

> My frustration that I have is that in many ways they can demonstrate through homework, through classroom participation, through the questions that they ask, even in conjunction with the homework assignment, what it is that they are learning. And if they are really, really proficient, they tend to be really good at that. When it comes to the actual taking of the test, it just seems that there is something in it that makes it— . . . the test does not necessarily test their knowledge of that particular item. (KD, Mathematics)

Teachers noted how much the assessment differs from classroom practice: "There are modifications made in the classroom. We reach out and differentiate the instruction and are able to accommodate. On a test, how can you do that?" (RV, Mathematics) One teacher considered the different strengths and weaknesses of one of her students:

> She achieved with honors in the skills category, but in the concepts and the problem solving, she was below in both. So I know she does . . . her writing was [a] weakness for her, and any time she had to write her name, that's where she struggled. She had a lot of gaps within her writing and her spelling. She was not a very strong speller. That could have contributed to her low scores with the problem solving and the concepts. (ER, Mathematics)

Both mathematics and special education teachers emphasized the reading challenges present in the mathematics assessment. Six mathematics and three special education teachers specifically mentioned the reading challenges their students face. One teacher noted, "That's a common place where students will not get credit for something that they may have knowledge of, but they don't really understand what the question is asking." (AE, SPED) One special education teacher noted that stronger readers do better on the mathematics assessment:

[1] All states have an alternate assessment for 1% of the population.

> The tests—when I think about who is really successful with it, are the readers. Students who are real readers can look at a test, and it can make real sense to them. It is words on a page. Other types of learners need to be more familiar. They need to work with what is familiar to them. The tests are presented in such a format that is not what they are doing in the classroom. (JS, SPED)

Others agreed that for weak readers, the format of the test becomes even more important:

> A lot of time kids don't understand what is written on the paper and what is expected of them. In the classroom, she [the mathematics teacher] not only explains it to them, but she shows them how to do it. You are giving the visual and the hands-on at the same time. On an assessment, you are given the paper, and you have to read it. Sometimes the kids have difficulty with reading comprehension. If it can't be read to them, they don't understand what is expected of them. (GG, SPED)

One format issue that was mentioned by a number of teachers is the "mixing up" of problem types:

> But the problem comes, again, if I were to give them a test that has different types of questions from different sections, they would have a hard time focusing on what they have and have not done. Because they will think "algebra for this section. And then they will have a hard time going back to fractions for this section. (JT, Mathematics)

For others, the physical changes have an impact: "Desks are put into rows so there is no cheating. The atmosphere is different. They walk into class and it is tense." (ST, Mathematics) The paper format limitations can also present a challenge:

> A lot of kids will take . . . if they have an addition problem, take some numbers from one problem and take other numbers from another problem because they have lost track of what they are working on. Because they have too much on a page. They also need more space to work out a problem on the same page. That visual tracking can slow some kids down. (GG, SPED)

The assessment context differs greatly from the classroom context when students are learning below the minimum level measured by the assessment. In these cases, students take the assessment, but it never measures any progress. One teacher commented:

> I mean, I don't think we've ever had kids that had no progress, but not necessarily a year's progress. It really depends on—but they are certainly making progress on their IEPs. And I think what has been frustrating is that we don't think the state testing measures their progress. Because we have kids that maybe have made progress in math or reading, but they still get a 200 or just a little over the 200 [the lowest possible score] on the math. (PT, SPED)

Thus, even with all of the attention paid to universal design in the development of large-scale assessments, teachers identify structural barriers for certain students. The reading level and test organization were the two barriers most often mentioned in these interviews, affecting students in both gaps 1 and 2.

Relationships: Teacher practices promote alternative demonstrations of learning

When teachers described their perceptions of why large-scale assessments might not measure student proficiency accurately, they talked about the differences between their classroom practices and the assessment. One teacher rattled off a list of the differences:

> There are tons of resources. There are tons of human interactions. There is [sic] varying paces and styles. There is [sic] ways for them to decode, and it is not all verbal. It is not all written. They get graded on it. There is something about it that affects their life. Absolutely everything that isn't on the testing situation, you can pour into the classroom. (CW, Mathematics)

Another teacher from the same school reiterated the difference between the classroom and assessment contexts:

> The classroom is never something on paper that you are handed, at least not in our math department. There is a lot of verbal. There is a lot of up and moving and doing. That is the exact opposite of a test. It is just you, words, and numbers. In terms of a learning style, it [the assessment] goes against everything you give kids. It is tough for them to say, "Wait a minute, in this we would be doing an activity, a model, we would be doing something that would get us going in that direction. I just have to grind it all out in my head." It is against everything we do with the kids. (RH, Mathematics)

Both teachers noted that human interaction and the pace and variety of learning opportunities are critical for some students.

In many classes, students have access to resources during classroom assessments. Said one teacher, "They are allowed to use manipulatives, their textbooks, and their binders for pretty much all of classwork to help them. I think they become dependent on them . . . So we teach them one way and then we test them another." (ST, Mathematics) Teachers commented how students in both gaps benefit from certain classroom styles:

> Those kids [in the gap] love putting things together, the hot-glue guns, the special problem solving, digging it apart. They love that, and they will do that for you. They love to know which buttons work on the calculator, and they love to make spreadsheets on the computers, and they do like all that concrete stuff that they can see . . . then you come to the test, and it is two paragraphs of words. It is no colors. There is nothing to fold. (CW, Mathematics)

Teachers gave many examples of ways that their positive classroom practices differ from the large-scale assessment. One teacher noted that students respond best to "video. Action. It is

really hard to make that bridge between the visual and the active and put it down on paper." (JS, SPED) Over and over again, teachers reiterated how their classroom practice teaches one way, and the assessment tests them another way.

Students learn and show their learning in relationships with supportive adults

Across all schools and all states, teachers observed that the individual relationship between student and adult is key to the learning process, and that when the assessment process breaks that relationship, student demonstrations of proficiency suffer:

> They're able to . . . like, in the classroom, they are able to have discussions and talk, ask questions, and talking back to me, and obviously during assessment, they are not able to do that. They have to sit and be silent, so there's a lot more verbalization and discussion within the classroom, which is obviously good for them. (EM, Mathematics)

One teacher shared his frustration with a student who can verbalize the process to solve a problem, but then cannot do the same problem on the assessment:

> I could sit down with this young man and say to him, "All right. Let's talk about number 3. What do you think you should do?" He could take a look at that: "Number 3 expects me to divide these two numbers into each other." "And how would you get the answer?" He says, "Well, I put these two numbers together and I'd come up with"—"What would the answer be?" "The answer would be seven." "What did you do when you took the test?" "I don't know." It is maddening. (KD, Mathematics)

This teacher used a relationship with the student to promote demonstration of learning. Another described particularly effective instructional assistants:

> I have wonderful [instructional assistants], absolutely wonderful. Just been my experience to have absolutely wonderful people. This is hard to put data on, even though I am a math person. What these people do, and then they get work out of these kids. They build the relationship almost first. They always have a smile on their face. They always have a positive attitude. Even when those kids come in with their negative demeanor, they, through their body language and . . . they don't do discipline . . . they do provide structure and security in a relationship. Then they can build. I don't know how you can get that into an exam. (SE, Mathematics)

This idea, that there are critical aspects of learning that "you can't get into an exam," was repeated across teachers. Another teacher described how relationships can help to mitigate other challenges, and argues that the assessment process doesn't include that critical element:

> It takes a long time to build up a relationship with a student that says, You can do this and I can help you and we'll work together . . . we have a very high population of students that come from homes that are not the best situation for these kids, so coming to school means they're safe. So it's not all about math, it's about making relationships, and I think

> that's part of what the test misses. It doesn't assess the fact that these kids are making relationships and social time and trying to make [it] through life. (RK, Mathematics)

Some students benefit greatly from human interaction, whether because they need the encouragement or because they actually need help figuring out how to start. Said one teacher, "When they are asked a complex question, if someone could start them, they could proceed. But when they can't be started, they will lose it." (PR, SPED) Another noted, "When I had him in the sixth grade and I tutored him in the seventh grade, he did really well, one on one. But when it came to taking any type of test, he froze, just because he knew he was being evaluated by it." (LM, Mathematics administer)

Some teachers focused especially on the way that the assessment experience is radically different from the way they teach in class. They find that their teaching pedagogy leads them to teach with many different styles to provide access to the curriculum in many different ways, depending on student learning styles, but the assessment is a single form of measurement that works well for just one type of student. Many other students, who thrive in alternative settings, don't "translate" their learning into the assessment context.

For other students, the long time period required to sit still becomes a challenge:

> They can't sit there for an hour and a half and concentrate on it and do one thing that long. They have to touch it and walk around it . . . I wish I could write an assessment that people would statistically trust, that would show them in their best light, which is interpersonal and interaction, where they are thinking and talking and trying and being positive. I have never seen an assessment that looks like that. (CW, Mathematics)

The lack of collaboration during the large-scale assessment is antithetical to some teachers' pedagogy. They feel like their students are thinking, "This is not how I normally do math. Normally the teacher is talking or I am working with a partner. I talk to somebody." (JS, SPED) Teachers argue that students who benefit from working together can struggle during an assessment: "The strengths that I see are that they typically work well together. They problem-solve as groups, while testing is individual." (JS, SPED) In addition, teachers describe the contradiction for students between classroom practice, which emphasizes collaboration, and the assessment, an individual exercise:

> They are always encouraged to work with someone in the classroom, during team time. I truly feel that problem solving requires you to think as a group—you don't usually do it independently. So talking through your homework and sharing what you know is really encouraged. (FL, Mathematics)

In class, students have interactions with teachers and collaboration with peers, neither of which is present during the assessment.

Students are able to receive clarifications to understand task demands

For many teachers, it is very frustrating to be prohibited from providing any of the scaffolding so commonly used in the classroom. One teacher noted, "We cannot read a question to a student,

we can't explain or interpret the problem for them and say, 'This is what's being asked here.' We're not allowed to do that. . . . Because you know that they could do it if they understood what they were supposed to be doing." (PC, Mathematics) The inability to give just a little nudge and help to "jumpstart" students frustrates teachers: "I think they are just kind of sucked dry when they see two pages of reading to do, where you can't help with anything. 'Could you just tell me what this one word is?' And you just can't—I think they, over the length of the test, they lose their motivation. (NB, SPED)

Another agreed:

> I think because [the assessment] is presented in a different way, a lot of time kids don't understand what is written on the paper and what is expected of them. In the classroom, she [the teacher] not only explains it to them, but she shows them how to do it. You are giving the visual and the hands-on at the same time. On an assessment, you are given the paper and you have to read it. Sometimes the kids have difficulty with reading comprehension. If it can't be read to them, they don't understand what is expected of them. (GG, SPED)

Teachers also talked about the link between relationships in class and test anxiety:

> In the classroom, they feel comfortable making mistakes. They are familiar with the teacher, and they are familiar with their classmates. If they get a wrong answer, it is okay. On the state assessment, when they think it is the . . . 'I must be a failure if I don't pass it. I can't get it wrong,' and maybe that goes back to the anxiety. I think the classroom is less stress. It is a comfortable learning environment. (ST, Mathematics)

The classroom setting allows teachers to provide reassurance and scaffolding for students, which helps students both to learn and to demonstrate their learning:

> They often can do well in a group, cooperative learning setting . . . when you actually make the connection to something they already know. Saying, "Oh, this is just like such and such. Remember you did this? The only way it's different is that . . ." When you verbalize the connection to previous learning, rather than assuming that they're going to make that connection. (AE, SPED)

This opportunity to verbalize the connection to previous learning doesn't exist during the assessment.

For another teacher, a critical issue is the inability to provide different options for assessment: "When I present a new topic, I try to present it using two or three different ways. Whether it's hands-on, or—I mean, I try to use as many manipulatives as I can to try to make that open, and then different ways of practicing." (JM, Mathematics)

Relationships are critical for students in both gaps. For students in gap 1, who do know the material, relationships can be the key to demonstrating their learning. For students in gap 2, who

are learning far below grade level, much of their learning can be based on one-to-one contact, making the isolated assessment a foreign experience.

Relevance: Student engagement depends on relevance to learning

A number of teachers mentioned the importance of having students be motivated to learn and motivated to take the assessment. Some teachers have noted that students who don't care about the assessment, or who do not find it relevant to their learning, are less likely to put out the effort needed to demonstrate their actual proficiency. This section looks at relevance in two ways: first, whether students think that the assessment is important to them, and second, what happens when the assessment content is not relevant to what students have learned in class.

Challenges with connecting large-scale assessment to students' daily lives

The most literal way that relevance is important to demonstrating proficiency, according to teachers, is that if students don't see the relevance, they won't put in the effort. One teacher noted:

> The lack of investment that the kids have. It means nothing to them. They never see it. They never know. There is no investment for them. It means nothing to them and [they are actually told], "it means nothing to you." They are told it does not affect their grade. The fact that they don't get the encouragement and feedback, and it offers nothing to them. All those things conspire to produce a situation they are not interested in doing and being engaged in. (CW, Mathematics)

If students choose not to engage in the assessment because they don't see its relevance, this presents a different challenge to schools. Should the assessment be changed to make it more relevant? What would that look like? Should schools put energy into trying to change students' attitudes? An administrator commented:

> Okay, I have one girl. She's very smart but doesn't see the relevance of taking the [assessment] . . . And what really happened was that she just didn't even want to be bothered with it and just filled in the blanks. Because [the students] are in a different place. (LM, Mathematics administrator)

It is not the task of assessment developers to motivate students to put effort into the large-scale assessment, but at the same time, this relationship between relevance and motivation is an area that concerns many teachers, and is, ultimately, an important question for all educators. External pressure to perform on an assessment can never be as effective as intrinsic motivation to learn.

Alignment of classroom teaching, curriculum, state standards, and teacher expectations

The second element of relevance has to do with the relevance of the large-scale assessment to the content students learn in class. At most of the schools that participated in the study, the teachers said that their curriculum is aligned to state standards. At one school, however, the teachers said

that the alignment has not yet taken place, and so teachers are on their own to try to cover the material that will be assessed: “I mean, it may be written someplace but nobody has—and I’m kind of in tune to things. It is not like we have had a workshop on it. It is not like we have really worked toward making that match-up.” (BN, Mathematics) If teachers are not matching their content to the standards being assessed, then the large-scale assessment will not measure what students have learned.

At other schools, though the alignment has been completed, teachers think that the standards are too high and that the students are not prepared to be successful at meeting them:

> I think if you took the top 5% and excluded them, 95% of them are coming in with skills that are lower than they should be. The GLEs say [the students] should be able to multiply through 15. I don’t know too many kids that can do that. I don’t know too many adults that can do that. Seven, 8, 9s are difficult. When you don’t have that solid foundation of basic math facts, nothing else is going to be easy. If you are worried about the calculation and not the concept, you are not going to get it. (ST, Mathematics)

Many of the mathematics teachers interviewed were quite open about the challenges they face in covering all the material in the standards: “There are gaps because we find that we cannot get through the amount of material in a year with the kids that we really need to.” (SE, Mathematics) One teacher explained:

> There were chunks that we didn’t get to. We didn’t spend a lot of time reviewing circumference. We didn’t spend a lot of time on these other pieces. And they either did a quick review or it was something that they had in sixth or seventh grade that we didn’t review. Or it might even have been something we had gotten to because they took the test in March, and we still have March, April, May, June. And we don’t get to some of that. (BW, Mathematics)

For some, the problem lies in the previous year’s teaching. A number of teachers talked about receiving students from more than one feeder elementary school, and finding that some schools tended to prepare their students more than others. Others argue that students’ different developmental trajectories mean that some are not ready for algebra.

> Just the difference between, you know, a blank space and a variable will deter kids who are not comfortable with that concept. But that is also a—I mean, we are supposedly believing that algebra is something that you need to be ready for. And if we throw it in and kids aren’t ready for it and they are being tested for it, it is not testing their math skills, it is testing their algebra skills. (SL, Mathematics)

Among the teachers interviewed, some emphasized the challenges that they face with their most low-performing students, while others argued that with enough effective teaching, every student can demonstrate proficiency on the large-scale assessment.

One teacher said about a low-performing student, “She will never, ever achieve, ever achieve. I don’t care how well I teach. I don’t care how well anyone else teaches. She is never going to

have the ability to achieve what these tests want her to achieve." (BW, Mathematics) Another talked about a "nice, good-hearted kid" who "picks a lilac on her way to school and gives it to a teacher," but who will probably not demonstrate proficiency on a mathematics assessment.

Other teachers take the opposite position. One teacher complained about offering only one algebra class in eighth grade, limiting the number of students who can take the course:

> We were limited to one class, and I've pushed for the last four or five years in saying, "I don't care if the kid didn't qualify as a seventh-grader for pre-algebra. But why are we going so slow with an eighth-grader? He is ready. Let's get him into a pre-algebra program now, even though it is not going to change what he does at the high school level. He should have some incentive to move. Keep pushing." (JT, Mathematics)

One teacher put forward a challenge:

> When it comes to the test, tell me, these are the items that are going to be on the test. And I can guarantee you, I can get every kid to that target, every single kid. And then it's about the clarity of it. And for some kids, just the stress of the whole thing shuts them down. Some kids just go and freak out. And it has nothing to do with the test, it has to do with stress. (AB, Science)

In a related survey of 56 seventh grade mathematics teachers in three of the participating states, the teachers were asked to report whether they covered each of the content standards and strands during seventh grade (the content to be assessed in the fall eighth grade assessment).[2] The results showed that only four of the teachers covered all the content in the 16 mathematics strands, and only 6 of the strands were reported as being covered by 75% or more of the teachers. This indicates that many students, not only those who are actually being taught below grade level, are not being taught the material that is being assessed.

The questionnaire asked teachers to describe their school's curriculum. Eighty percent of teachers reported that their school districts had a required program or curriculum for seventh grade math. Alignment of the districts' math programs to the grade-level expectations (GLEs) was reported as complete by 24% of the teachers and as more than half-completed by 47% of the teachers. Another 25% of the teachers reported that their districts had started the alignment process. GLEs for Numbers & Operations and Functions & Algebra were the most completely aligned to district programs (listed by 85% and 78% of teachers, respectively). Data, Statistics, & Probability was least aligned (66%). Coverage of the seventh grade math GLEs in 2004–2005 varied by the four strands. Numbers & Operations was covered most thoroughly, with most variability in the coverage of Functions & Algebra. Teachers felt that Functions & Algebra was most difficult and Numbers & Operations was least difficult for seventh grade students.

Relevance of content is multi-faceted. One element is the need to align classroom teaching with standards, which is an instruction-focused task. Another element is whether students are being taught the grade-level material. On the surface, this is an instruction-focused task, but it also

[2] The survey was conducted within the same U.S. Department of Education Grant CFDA #84.368, and was developed and analyzed by Sue Bechard of Measured Progress.

must be looked at in light of the two assessment gaps. If students in gap 2 are not being taught the material, this has implications for the assessment. And if students in gap 1, who generally perform on grade level, are also not being taught the material, this may require a shift in instruction, or a re-examination of the standards. A third element is the power of teacher expectations. As shown here, some teachers are quick to identify students incapable of learning, while others are adamant that all students can learn. Some of these elements may best be addressed through instruction, but for students in gap 2 in particular, the difference between classroom teaching and the large-scale assessment becomes a critical issue.

Conclusion

In this study, teachers described differences between the classroom context and the assessment context in three ways: the structure of large-scale assessment creates barriers to demonstrating proficiency; teacher practices promote student demonstration of learning in ways that large-scale assessments do not; and students respond better to assessments that are relevant to their learning. Some teachers advocated making the assessment more similar to the classroom context, focusing in particular on the need for some students to interact with another person as part of the assessment process, and the need for other students to have an opportunity to be assessed using alternative means, such as projects or hands-on demonstrations.

In both gaps 1 and 2, this study could not definitively delineate when the gap is truly an assessment gap that should be addressed through assessment changes, and when the gap is instruction-based. Teachers said that in many cases, the mathematics assessment does assess students' levels effectively, and that those students who do not demonstrate proficiency do not, in fact, know the material. Thus, the assessment is doing what it is supposed to do. This leaves us with the unanswered question of how much of the gap lies with the assessment and should be addressed by changes in the assessment, and how much lies with instruction. The teachers in this study, while arguing that the large-scale assessment fails to account for many facets of student progress, also tended to agree that the mathematics assessment is a good measure of the standards it claims to assess. This study was only able to subjectively identify those students who would benefit from a different assessment rather than from different, or more, instruction.

This study also indicates that different stakeholders may have different understandings of the purpose of assessments and the definition of proficiency. Teachers interviewed for this study described ways that differences between the assessment context and the classroom context negatively affect students' ability to demonstrate their proficiency. However, depending on the definition of proficiency, large-scale assessment may be an appropriate measurement tool. If proficiency definitions include the ability to transfer skills and to solve mathematics problems in different contexts, then students should be able to transfer the proficiency they demonstrate in the classroom context to a large-scale assessment context. Similarly, if proficiency definitions include the ability to work independently, or to focus on a task for 45 minutes, then a large-scale assessment can be a good way to measure that. It may be that when considering the ways in which large-scale assessments can be made more accessible to all students, two other areas should also be included: the degree to which students are receiving instruction in the areas being assessed, and the definition of proficiency. If those aspects that are part of large-scale assessment (transferring skills from one context to another, working independently, and maintaining focus for an extended period) are also part of the definition of proficiency and are somehow measured,

they would become part of the skills set that teachers need to encourage in their students. This also presents a challenge to assessment developers: to develop an assessment that can separate the constructs of mathematics skills from the above-mentioned constructs of skill transfer or working independently.

Of the three areas identified by teachers—structure, relationships, and relevance—only issues related to the assessment structure can be addressed within the current standardized assessment framework. Teachers reiterated that some students need to have human interaction in order to show what they know and can do, but incorporating relationships into the standardized assessment framework presents formidable challenges. And the issue of relevance (for students in gap 1, the relevance or lack of relevance of the test-taking process, and for students in gap 2, the lack of relevance of an assessment that does not measure what they are learning) raises a series of questions about the larger framework of content and skill standards: What should be done if students choose not to engage in learning and assessment because they do not see it as relevant to their lives? How can students who are progressing in their learning, but are still far below grade level, be assessed for their progress without lowering the expectation that they perform at grade level?

Further research is needed to confirm or disprove the framework presented in this paper, in particular, whether the three differences between the assessment context and the classroom context are also identified as the most important issues by other stakeholders (administrators, students, parents), and whether it is possible to address these issues (or others that emerge) in large-scale assessments. Further research is needed to determine how to accurately assess those students who need “something to fold” to show their knowledge and skills. Policy work needs to be done to address differential understandings of the purpose of large-scale assessment and the role of proficiency definitions, and to consider the role of instruction in the assessment gaps.

References

Almond, P., Quenemoen, R., Olsen, K., & Thurlow, M. (2000). *Gray Areas of Assessment Systems.* NCEO Synthesis Report No. 32. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Bolt, S. E., & Thurlow, M. (2004). Five of the Most Frequently Allowed Testing Accommodations in State Policy. *Remedial and Special Education, 25*(3), 141-152.

Edgemon, E. A., Jablonski, B. R., & Lloyd, J. W. (2006). Large-Scale Assessments: A Teacher's Guide to Making Decisions about Accommodations. *Teaching Exceptional Children, 38*(3), 6-11.

Quenemoen, R., Thompson, S., Thurlow, M., & Olsen, K. (1999). *Forum on Alternate Assessment and "Gray Area" Assessment.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the Development and Review of Universally Designed Assessments.* NCEO Technical Report No. 42. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Universally Designed Assessments. (2005). Retrieved May 5, 2006, from http://education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_topic.htm

Wiener, D. (2006). *Alternate Assessments Measured Against Grade-level Achievement Standards: The MA "Competency Portfolio".* NCEO Synthesis Report No. 59. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., Thurlow, M., Bielinski, J., House, A., Moody, M., & Haigh, J. (2001). The Relationship Between Instructional and Assessment Accommodations in an Inclusive State Accountability System. *Journal of Learning Disabilities, 34*(3), 212-220.

Appendix A. Interview protocol for initial teacher interviews: Looking at assessment gaps

Think of the students in your school who did not reach proficiency on the eighth grade math state assessment.

1. Do these students fall into any particular categories? If so, what are they?

(If students who don't reach proficiency but who demonstrate proficiency in class are mentioned, direct participants to this group. If these students aren't mentioned, mention them.)[*]

Think of students in your school whose class performance indicates that they really do have the knowledge and skills to reach proficiency on the state test, but whose assessment results are below proficiency.

2. Can you describe these students for me?

    Gender

    Age

    Ethnic background

    SES

    Second language issues

    Learning issues/SPED

    Learning style, strengths, weaknesses

    Academic history—classroom settings, curriculum they were exposed to

    Opportunities to learn

    Family history/background

    Test anxiety/previous testing history

    Attitude toward math, math phobia

[*] We initially targeted only students in gap 1, but teachers consistently described students who fit into a different assessment gap, which became our gap 2. Teachers answered the questions for both gap 1 and gap 2 students.

Attitude toward giving up early

Classroom settings—any special help provided for math or for test preparation

3. On what classroom evidence do you base your belief that this student should have demonstrated proficiency on the state assessment?

   How did you collect this evidence?

   To what state performance level (proficient, beyond proficient, etc.) does this student's classroom performance correspond?

   What kinds of grades does this student get in math classes? How did this student perform in math in school? What are their strengths in math?

   To what kinds of instruction or activities do this student respond best?

   To what kinds of assessment types do this student respond best?

   In your opinion, what about the state assessment makes it difficult for this student to demonstrate what they know and can do?

   What is it about classroom instruction that enables this student to show what they know and can do?

   What would be a better way of measuring what this student know and can do?