Reclaiming Assessment Through Accountability That Is "Just Right"

ELIZABETH GRAUE
ERICA JOHNSON

University of Wisconsin-Madison

Background: *This article builds on three years of qualitative research on Wisconsin's Student Achievement Guarantee in Education (SAGE) program, a class size reduction policy in Wisconsin.*

Objective: *In this article, we take a practice-oriented perspective on assessment, examining how assessments in schools that participated in a class size reduction program intersected with forces of accountability. The goal of this article is to broaden the understanding of what it means for schools and teachers to be held accountable for student learning and to discuss how different accountability frameworks affect instructional practices in classrooms.*

Setting: *The research took place in nine elementary schools across South and Central Wisconsin.*

Research design: *Data for the qualitative case studies were generated through multiple methods, including ethnographic observations, interviews, administration of the Classroom Assessment Scoring System (CLASS), document and artifact collection, and analyses of school-level standardized test scores.*

Results: *The current political and educational context oriented our focus to an intersection of issues: the implementation of class size reduction, instructional practice, assessment, and accountability. We identify three aspects of assessment practices that affect this intersection: alignment, audience, and action.*

Conclusions: *We found that coherent and collaborative assessment practices were more likely to take place in schools where there were explicit connections through assessments to varied communities of interest: district, school, teachers, students, and families. In supportive assessment systems, teachers had tools that they understood and that they could use to improve their practice to meet the needs of their students. In contrast, assessment in lower quality classrooms took place in disjointed systems that focused primarily on summative rather than formative assessment. A focus on accountability without attention to the quality of instruction and the quality of assessment resources is inherently flawed.*

Teachers College Record Volume 113, Number 8, August 2011, pp. 1827–1862

0161-4681

> Ultimately, accountability is not only about measuring student learning but actually improving it. Consequently, genuine accountability involves supporting changes in teaching and schooling that can heighten the probability that students meet standards. Unless school districts undertake systemic reforms in how they hire, retain, prepare, and support teachers and develop high quality teaching, the chances that all students will have the chance to meet new high standards are slight. (Darling-Hammond, 2004, p. 1078)

Despite a broader history and potential implications, in the United States, the term *accountability* has become synonymous with the high-stakes standardized tests required by NCLB. Although accountability is often depicted as a conservative force designed to dictate teacher action, it can provide an intentionality that ultimately makes teaching more effective. The goal of this article is to broaden the understanding of what it means for schools and teachers to be held accountable for student learning and to discuss how different approaches to accountability affect instructional practices in classrooms.

In exploring the links that can be made among assessment, instruction, and learning, our work takes a practice-oriented perspective on assessment. We build our case for a practice perspective from the recent discussion of validity by Moss, Girard, and Haniford (2006):

> Educational assessment should be able to support these professionals in developing interpretations, decisions, and actions that enhance students' learning. Validity refers to the soundness of those interpretations, decisions, or actions. A validity theory provides guidance about what it means to say that an interpretation, decision, or action is more or less sound; about the sorts of evidence, reasoning, and criteria by which soundness might be judged; and about how to develop more sound interpretations, decisions, and actions. (p. 109)

For us, placing assessment within educational decision-making and teaching is essential if assessment is going to have a strong relationship to practice. Based on case studies of practice in nine Wisconsin elementary schools participating in a class size reduction program, we examine how stakeholders took up the call of accountability and worked to raise test scores, enhance teacher professional knowledge for instruction, and increase their knowledge of students.

Links between class size initiatives and assessment have been posited by a number of scholars. Ease of assessment and improved accountability are two aspects of the theory of action that undergird the implementation of class size reduction reforms. By reducing the number of students in a class, teachers are thought to have more opportunities for formative and summative assessment, which provides information for more targeted instruction, resulting in increases in student achievement (Biddle & Berliner, 2002; Blatchford, 2003; Grissmer, 1999; Odden & Archibald, 2000). For young students, assessment and specific teaching of social and emotional content are viewed as a foundation for socialization into the practice of schooling (Finn, Pannozzo, & Achilles, 2003). Complementing prior research, we identified class size reduction, assessment, and accountability as themes in our fieldwork. To ground reading of the case studies presented in this article, we turn first to the literatures on two distinct approaches to assessment: standards-based accountability and instructional assessment.

REVIEW OF RELATED RESEARCH

The current context for education is quite often identified with the No Child Left Behind Act of 2001 (NCLB, 2002). Its use of high-stakes standardized assessments as the primary measure of student outcomes and school accountability has altered how researchers and practitioners alike characterize their work. However, it is important to recognize that NCLB is part of a historical thread that pushes assessment to the forefront of education debates. In our review of the related research, we explore two aspects of assessment that are often contrasted in the literature on reform. We employ a distinction coined by Paul Black and Dylan Wiliam (1998)—one that focuses on the purposes of assessment and ultimately the audiences they are meant to inform: assessment *of* learning versus assessment *for* learning. We explore how they have been connected in policy and practice and how they enhance and constrain each other, and set the groundwork for our analysis of their enactment in our study.

ASSESSMENT OF LEARNING

History. Scholars in education assessment, measurement, policy, and curriculum have documented the burgeoning test-based economy and its effects on teaching and learning (see for example, Koretz, 2008; Madaus, Russell, & Higgins, 2009; Nichols & Berliner, 2007; Ryan & Shepard, 2008). One could look back hundreds of years for links between teaching and testing, but one reasonable marker comes in a more recent historical era: A series of policies and reforms in the 1960s used program evaluation as a tool for justification of change, with data collection a key component of program design. Some reforms used proximal measures directly related to program intents (e.g., Title I of the Elementary and Secondary Education Act of 1965, which required funded schools to provide data on student outcomes), whereas others were framed with broader strokes to provide information on the general health of the education system (e.g., National Assessment of Educational Progress). These *reform-driven assessments* were low stakes and provided general information.

A shift came in the 1970s, when testing was suggested as a way to check whether students had learned baseline content expected for promotion or graduation. Tests were used to direct students' and teachers' attention to minimal competencies required to move from level to level in schooling or to graduate. These *assessment-driven reforms* were connected to serious consequences, making the test a critical tool in monitoring quality (Shepard, 2008). It is the shift to assessments leading education policies that most clearly resonates in today's high-stakes context.

A Nation at Risk (Gardner, 1983) prompted policy makers to turn away from minimal standards and design programs that aimed high by setting standards that would push for "excellence." Among the strategies used in these reforms were tests that drove instruction. These reforms were premised on the idea that to be most effective, practice should be guided by some kind of measure. Tests became the measure of efficacy. Faith was placed not only in the idea that tests were useful for and used to orient teaching and learning, but also in the idea that tests could in fact measure outcomes that represented that learning.

Standards-based reform and accountability. A clear definition of this approach is stated in a National Research Council (1999) report entitled *Testing, Teaching, and Learning:*

> Generally, the idea of standards-based reform states that, if states set high standards for student performance, develop assessments that measure student performance against the standards, give schools the flexibility they need to change curriculum, instruction, and school organization to enable their students to meet the standards, and hold schools strictly accountable for meeting performance standards, then student achievement will rise. (p. 15)

The theory of action inherent in this model is simple but not easy. Standards-based accountability relies on two related elements: alignment and capacity building. Alignment ensures coherence between the curriculum and the accountability system—but alignment is insufficient. Building capacity at the classroom and institutional levels ensures that staff have the skills, knowledge, and resources necessary to implement an aligned curriculum (Carnoy & Loeb, 2002).

In some systems, accountability relies on intrinsic motivation; through the mere act of recognizing standards and outcomes, both students and teachers will be motivated to work harder and more efficiently. Here, the system is seen as low stakes. In other systems, accountability is extrinsic, focused on incentives and sanctions connected to student outcomes. The system in this case has high stakes connected to assessments. Finally, the theory of action assumes "that barriers to improvement have lower strength than the desire to achieve goals and that there are clear and powerful incentives for powerful actions" (Baker & Linn, 2004, p. 48). This last point represents a recognition that implementation of standards-based accountability is dependent on systemic forces and that barriers exist to success.

NCLB. As standards-based accountability was implemented across the nation, tests overshadowed standards as the primary lever for reform (Carnoy & Loeb, 2002). Standards-based accountability was given a substantial nudge with the passage of NCLB, which heightened attention to test-based accountability at the federal level. The standards component of the reform was designed to provide clarity of purpose for districts and schools, with a shift from an individual entrepreneurial model for teachers and schools to a systems approach that made more intentional the workings of curriculum among the varied levels of schooling: classroom, grade level, content area, school, district, and state. As it became the favored mode of instructional improvement, researchers studied the effects of this reform on student achievement, policy implementation, and teacher and student experiences. We briefly review that research next.

When researchers focused on the classroom, they found that teachers change their instruction as a result of accountability systems, with external tests shaping both the focus and the nature of instruction (Herman, 2004; Nichols & Berliner, 2007; Rothstein, 2008; Valli & Chambliss, 2007). Tested content in literacy and mathematics takes the majority of classroom time, reducing time spent on science and social studies (Diamond & Spillane, 2004; Shepard & Daugherty, 1991; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1991; Wills & Sandholtz, 2009). Test-like formats become dominant in general instruction and in actual test preparation activities (Herman). Koretz, McCaffrey, and Hamilton (2001) identified a range of teacher responses to high-stakes testing, including giving more instructional time, covering more material, working harder, cheating, reallocating instructional time, aligning instruction to standards, and focusing on specific aspects of tests. Although accountability systems have the potential to enrich learning, in practice, they also narrow the curriculum. The translation of standards through assessments results in impoverished content and formats.

Test scores as accountability and data. Standards-based accountability is a systemic approach, linking classrooms to schools to districts to state systems. Individual measures of students are aggregated to allow evaluation of various units, including classrooms, grade levels, schools, districts, and states. For this reason, it has been important to examine the effects of accountability programs in these nested contexts. When instruction is connected to broader systemic issues, responses to test-based accountability are related to the school's accountability status (Diamond & Spillane, 2004; Pedulla et al., 2003). In low-performing schools, instruction is targeted at raising test scores, often by focusing on groups of students at the borders of proficiency. In high-performing schools, instruction is designed to raise the achievement of all students. This pattern is repeated in classes designed to serve students with comparatively lower or higher levels of achievement (Valli & Chambliss, 2007). One particularly pernicious practice is a focus on the "bubble kids," those students hovering just below the proficiency level. With sanctions that focus on students who are not proficient, schools game the system, targeting instructional interventions with the highest probability to move into the proficient category (Koretz, 2008). Students not in this group are essentially ignored or given minimal attention. These differences result in a systemic widening of the gap in education quality, particularly because low-performing schools and classrooms tend to serve children of poverty and of color.

The structure of standards-based programs includes setting standards and developing assessment programs related to the standards. A critical element of most systems (and instantiated in NCLB) is the delay of standardized testing until Grade 3. The logic of this method is that young children are notoriously difficult to test and that any results derived from testing students younger than 8 years of age would be highly prone to error (Shepard, Kagan, & Wurtz, 1998). Accountability systems reach into contexts for younger students, however, because school personnel use standards and tests geared for students in Grade 3 to back map benchmarks to the primary grades and even pre-K (Brown, 2007; Goldstein, 2008). This is relevant for our study because we worked in kindergarten through Grade 3 classrooms.

How valid are judgments made from accountability systems? Much of the faith the public puts in them is based on the systems' presumed objective nature. The process of constructing the system, including setting standards and choosing proficiency levels, is a very human process, filled with political negotiations and technical discussions (Ellwein & Glass, 1991; Smith, 2003). Current analytical strategies make it difficult to link accountability indicators to the efficacy of a single teacher, curriculum, or school. In addition, increases in test scores follow repeated use of an accountability test; this increase disappears if a new tool is used (Linn, Graue, & Sanders, 1990). Paired with evidence that many educators are moved to narrow their curriculum to mirror the test content and formats, it is clear that interpretations of high-stakes accountability indicators are vexed at best.

Summary. Assessments are a key attribute of standards-based accountability. Advocates argue that the power of the alignment process and the motivation of the potential for negative consequences together provide the focus to improve education (Smith & O'Day, 1991). Critics note that high-stakes testing often goes beyond healthy focus and results in narrowing of the curriculum and tactics that are aimed to game the system. These instructional practices make interpretations of the test results from accountability programs difficult to interpret. At the same time that standards-based accountability and their associated tests have taken center stage in education reform, another form of assessment, one explicitly designed to inform instruction, has captured the interest of assessment and content scholars as well as teachers. We turn to the literature on assessment for learning next.

ASSESSMENT FOR LEARNING

Classroom assessments, developed by teachers to track student learning, were often seen as so idiosyncratic and lacking rigor that they had little use beyond the teacher's desk. The past 20 years have seen a growing interest in the potential of assessments close to teaching practice, suggesting that teaching and learning can be enriched when assessments provide high-quality information to inform teachers and students (Black & Wiliam, 1998; Natriello, 1987; Shepard, 2005).

Definition. Assessment for instruction is a concept well known in the teaching community, but it goes by many names. Variously called formative (contrasting it with summative), classroom (signaling where it happens and where its primary audience practices), instructional (focusing on the practice it hopes to influence), and assessment for learning (contrasted to assessment of learning to show process vs. product intentions), assessment that is used to inform instruction is critical to good teaching and learning.

Researchers who have reviewed the literature on classroom assessment have taken different approaches to the task. Natriello (1987) identified four purposes of evaluation and described effects on students according to evaluation purpose, task resolution, criteria clarity, level of standard demand, standard referent, evaluation frequency, differentiation, and affective nature of feedback. He recommended that future research, if it was to inform practice, would need to recognize the different assessment purposes that exist in education.

Black and Wiliam narrowed the field in their 1998 review and focused specifically on classroom formative assessment. When Black and Wiliam reviewed the literature on classroom assessment, they found that when well designed and appropriately implemented, assessment for learning has strong effects on student achievement across content areas, producing effect sizes between 0.4 and 0.7 (Black & Wiliam, 1998). Equally important was that good classroom assessment was particularly effective in supporting the learning of students who typically achieved at lower levels.

Process. Effective assessment for learning is defined as "the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go, and how best to get there" (Assessment Reform Group,[1] 2002, p. 2). A critical attribute in this definition is that the actors include both teachers and students, with data informing decision-making and action. The Assessment Reform Group (1999) outlined a number of attributes of effective assessment for learning. First, assessments need to be designed to provide usable feedback that helps individuals close the gap between a current state and desired outcomes. Assessment is not something done *to* students, but it is an act that requires participation and provides information for both parties. In fact, self-assessment is seen as a critical aspect of effective assessment practices. Beyond the assessment act, instruction should inform and produce changes in instruction. High-quality assessment for learning is designed to recognize the profound influence assessment has on the motivation and self-esteem of pupils, both of which are crucial influences on learning. These characteristics, motivated by the belief that all students can learn and achieve, make assessment a constitutive part of learning, envisioning an active role for students in their own learning (Assessment Reform Group, 1999).

Under NCLB. In a context where external testing exerts strong influence on instruction and where teacher preparation in instructionally relevant assessment practices lags behind their needs (Shepard, Hammerness, Darling-Hammond, & Rust, 2005), the role of professional development has been addressed by researchers through studies that have practitioners partner with assessment researchers. In this work, researchers work with groups of educators to develop assessment capacity with a group of teachers through discussion of assessment tasks, content, and expected developmental patterns. The capacity typically is built on two levels—individuals learn about assessment and connecting it to their instructional knowledge, and the group develops a social knowledgebase that provides support (Shepard, 2001; Stiggins, 2005). Teacher assessment practice appears most effectively facilitated by active participation in the development of linkages to curriculum and implementation of assessments, specific professional support on assessment strategies (Black & Wiliam, 2004; Forster & Masters, 2004), and tools for the assessment process (Frederiksen & White, 2004). By making assessment a critical part of instruction and reinforcing links to decisions made on the basis of assessment data, assessment for learning is brought into the core of teacher activity.

Summary. The literature on assessment of learning has primarily focused on the effects of high-stakes testing on teaching practices and education systems. The research on assessment for learning has described how this type of in vivo assessment enhances instructional practice by making actions informed by systematically generated information. What is missing is examination of how accountability programs, including high-stakes assessments, shape the practice of assessment. The work described in this article brings together these scholarly discussions by examining how standards-based accountability reforms are related to the assessment practices in a group of case study schools. This approach is important because although both standards and assessments have incredible power to enhance teaching and learning, some combinations of the two are more conducive to achievement than others.

The implementation of a class size reduction program provides a fertile context for studying these connections given assessment's relation to the theory of action thought to justify smaller classes.

METHODS

This article comes out of a multiyear, multimethod study of Wisconsin's Student Achievement Guarantee in Education (SAGE) program. SAGE is a state-supported class size reduction (CSR) program that provides funding to districts to limit class sizes to 15 students and one teacher in Grades K–3 in almost 500 Wisconsin schools. Since 2004, our research team has followed nine SAGE schools in six districts. The diverse sample was purposefully sampled to include schools representing a range of poverty, urban, rural, and semiurban locations, and student achievement. See Table 1 for school demographic characteristics. Sampling for the initial study was based on data from 2003.[2]

Table 1. Schools 2006–2007

School	Bethany	West Canton	McMahon	Earhart	Calloway	Montford	Farmington	Wellstone Blvd.	Gallows
Geography	Urban	Rural	Semiurban	Semiurban	Urban	Rural	Rural	Urban	Urban
District	Mallard	West Canton	Bellamy	Maxwell	Mallard	Walton River	Farmington	Mallard	Mallard
2006 Enrollment	487	306	233	242	200	500	442	337	599
% Asian	6.4	.7	2.1	20.2	5.0	8.8	.5	6.5	.8
% Native Amer.	.4	0	.4	1.2	2.7	2.0	1.8	1.5	.2
% Black	80.3	1.3	31.8	31	13.3	3.8	.8	78.6	64.9
% Latino	1.8	1.6	21.9	14	44	2.2	5.9	6.8	19.0
% White	11.1	96.4	43.8	33.5	35	83.2	91	6.5	15
% Eng. lang. learners	1.4	0	13.7	39.9	1.9	8.4	.2	12.5	1.2
% Spec. education	14.1	13	26.7	10.1	9.4	15.4	24	13.8	36.7
% Free/red lunch	84	37.6	68.2	66.1	78.7	64.4	58.4	95.3	88.8
% Gr 3 reading prof/advanced	92	96	81	68	74	69	76	57	47
% Gr 3 math prof/adv	76	92	53	52	49	80	76	50	42
Classroom configurations	15:1 30:2	15:1	15:1	15:1	15:1	15:1 30:2 Core SAGE	15:1 30:2 Core SAGE	15:1	15:1
Performance relative to expectations[1]	3.15	1.62	4.82	3.46	5.93	-1.04	1.24	-8.65	-5.24
CLASS ratings by school	5.37	5.83	4.63	5.94	4.49	5.46	5.10	4.37	4.74

[1] School performance estimates combine three years of data from each school to provide an estimate of the expected percent of students proficient or advanced in Grade 3 reading, controlling for student characteristics and average teacher experience and training. Schools with a performance estimate of around zero are at expectations, given the population of students who took the tests and teacher characteristics. Schools with negative performance estimates are doing worse than expected, and schools with positive estimates better than expected.

Between 2004 and 2007, the study design alternated between in-depth case studies of practice and follow-up studies. In 2004–2005, we generated data through eight half-day observations in 27 classrooms (one kindergarten, one first-grade, and either a second- or a third- grade classroom in each school), standardized environment descriptions, collection of artifacts, and multiple interviews with teachers, principals, students, district administrators, and families. In 2005–2006, we revisited the nine schools to interview participants to gain more understanding of their practice.

In the 2006–2007 school year, research explored what we thought of as “best practice.” We selected three schools (of our original nine) for in-depth study because they appeared to represent high levels of implementation of the SAGE program and higher levels of student achievement. Through analysis of prior years’ standardized test score data and case study data, we determined that Calloway (urban), Earhart (semiurban), and Montford (rural) had recently improved student achievement and seemed to have in place reforms that were changing the culture at the school. In our focus schools, we returned to a kindergarten, first-grade, and either second- or third-grade classroom (the same participants as those in our initial sample when possible) to collect a diverse set of data.

In each of the case study schools, we completed seven half-day observations of instructional practice in each classroom. The observations in 2006–2007 focused on teacher–student and student–student interactions at varying levels, attempting to document the array of instructional practices. During a typical 3- to 4-hour observation period, a researcher recorded field notes with a laptop, moving around the classroom as needed. These notes were expanded upon leaving the field site, typically that evening or the next day. One observation consisted of videotaping a set of typical lessons. During these sessions, a professional videographer accompanied the researcher and set up two cameras in the room.[3] Across these various observations, we amassed a minimum of 25 hours per classroom.

During the year, we interviewed each principal twice and each observed teacher three times.[4] For many of these educators, this was the third year of interviews for the project, and they had a level of comfort with the process that made conversations rich. Interviews lasted between 40 and 120 minutes. Participants chose the interview locations and times, typically either during or immediately following the school day. Interview protocols were semistructured, meaning that the interviewer could ask follow-up questions for clarification as desired. All interviews were digitally recorded and transcribed. The interviews focused on the school and classroom practices related to SAGE’s four elements of practice (articulated next). During observations and interviews, we also collected a variety of relevant documents and artifacts in each of our case study schools, including worksheets, curriculum guides, and school and district reports.

We also collected limited data at the remaining six schools from the initial sample, visiting three classrooms (one kindergarten, one first grade, and either a second or a third grade) and conducting one interview with each school's principal and participating teacher.

In all classrooms in the nine-school sample, we administered a standardized observation known as the Classroom Assessment Scoring System (CLASS), which provides a common metric for understanding classroom quality (Pianta, La Paro, & Hamre, 2006, 2008). CLASS has a strong empirical record, used in more than 3,000 classrooms in the United States. Table 2 provides a description of CLASS domains, dimensions, and behavioral markers.

Table 2. CLASS Domains and Dimensions

Emotional Support	Classroom Organization	Instructional Support
Positive climate Enjoyment & emotional connections between teachers and students and quality of peer interactions	*Behavior management* How teachers monitor, prevent, and redirect behavior	*Concept development* How teachers promote higher order thinking and problem-solving
Negative climate Teachers and/or student negativity (anger, hostility, aggressions)	*Productivity* Effectiveness of teacher management and classroom routines to maximize learning time	*Quality of feedback* How teachers extend student learning through responses
Regard for student perspectives Teachers' interactions with students and how activities place emphasis on students	*Instructional learning formats* How teachers engage students in and facilitate activities	*Language modeling* How teachers facilitate and encourage student language
Teacher sensitivity Teacher responsiveness to individual needs		

Internal consistency of the three CLASS domains is moderate to high (α =.76–.95). Reliability across cycles was moderate to high (.68–.97), and CLASS scores are highly stable across days (Pianta et al., 2008). Although all research team members completed in-depth training and were certified CLASS coders,[5] one team member completed all CLASS observations in this data set. In each classroom, she observed and coded classroom practice across a minimum of four 30-minute cycles of instruction with CLASS.[6] After each observation cycle, the observer derived a rating of 1–7 on each of the CLASS dimensions. These observations also generated brief field notes for analysis. Ratings are categorized at three levels: 1–2 = low; 3–5 = midrange; 6–7 = high. Average scores were tabulated across four cycles, and CLASS dimension and domain scores were calculated by both classroom and school.

Data collection and analysis was an iterative process, guided by the interests and assumptions that shaped the study design and the issues that came up in fieldwork. Supported by the qualitative research software NVivo, our analysis began with a shared set of codes. We provide a list of codes (or *nodes* in NVivo) and examples of subcodes in Table 3.

Table 3. Examples of Codes and Subcodes Used for Analysis in NVivo

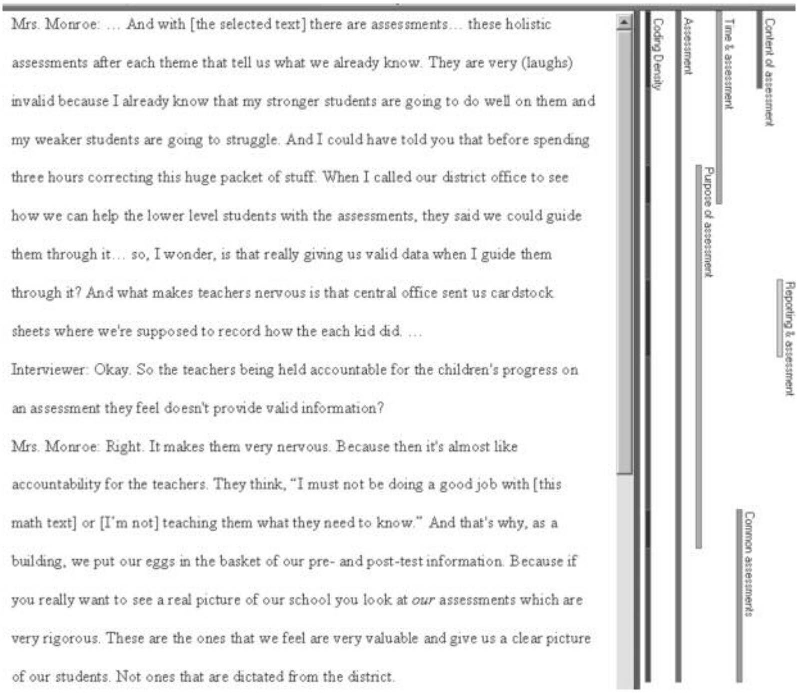
	Child Nodes (level 1)	Examples of Child Nodes (level 2)
CLASS	Emotional support	positive climate, negative climate, teacher sensitivity, regard for student perspectives
	Classroom organization	behavior management, instructional learning formats, productivity
	Instructional support	concept development, quality of feedback, language modeling
Class size reduction	Adults in classroom	one, two, three, four or more
	Grouping	flexible, free choice, individual, small, whole
	Numbers	15 or less, 16 to 20, 21 to 30, more than 30
	SAGE-related	justification for SAGE, SAGE research
	Space	space limitations, unconventional uses
Curriculum and instruction	Assessment	documentation of, observation as, purpose of, audience for, time and, content of
	Subject areas	arts, literacy, play, science, social skills, social studies, math, technology, health
	Curricula & programs	Direction Instruction, Reading First, Houghton Mifflin, Accelerated Reader, Marzano Framework, Responsive Teaching, Professional Learning Community
	Instruction	best practice, comprehension, conferencing, differentiation, ELL, games, homework, materials, morning routine, scheduling, time
Lighted schoolhouse	Home, family, life	comments on quality, assumptions
	Home-school communications	homework, notes, announcements, phone calls, visits
	Family participation	in classroom, at school, at home, formats of participation
Overview and context	School context	building and school changes, community, enrollment, health care, local culture, office interaction, SES issues
	External support	funding, state support, district support, community support

Table 3. Examples of Codes and Subcodes Used for Analysis in NVivo

Professional development	PD activities	goals, challenges, resources, funding, successes, results
	Planning time	uses of planning time, common planning time
	Collaboration	coordinated by admin, coordinated by teachers
School personnel	Administration	changes, support, leadership, principal, identity, shared leadership
	Staffing	assistant time, looping, unpaid classroom volunteers
	Teachers	expectations, autonomy, communication, mentoring, identity, leadership, union, interests, perceptions, quality
	Students	attendance, expectations, kid interaction, playful, resistance, independence, needs, mobility

As our work progressed, the team met periodically to share observations, interviews, and emerging ideas of issues in the field. Subsequent fieldwork reflected these discussions. We shared memos (Graue & Walsh, 1998) that detailed second-order analysis that linked coded data with cross-cutting themes. We shared much of our emerging analysis with our

Figure 1. First-pass analysis in NVivo



Mrs. Monroe: … And with [the selected text] there are assessments… these holistic assessments after each theme that tell us what we already know. They are very (laughs) invalid because I already know that my stronger students are going to do well on them and my weaker students are going to struggle. And I could have told you that before spending three hours correcting this huge packet of stuff. When I called our district office to see how we can help the lower level students with the assessments, they said we could guide them through it… so, I wonder, is that really giving us valid data when I guide them through it? And what makes teachers nervous is that central office sent us cardstock sheets where we're supposed to record how the each kid did. …

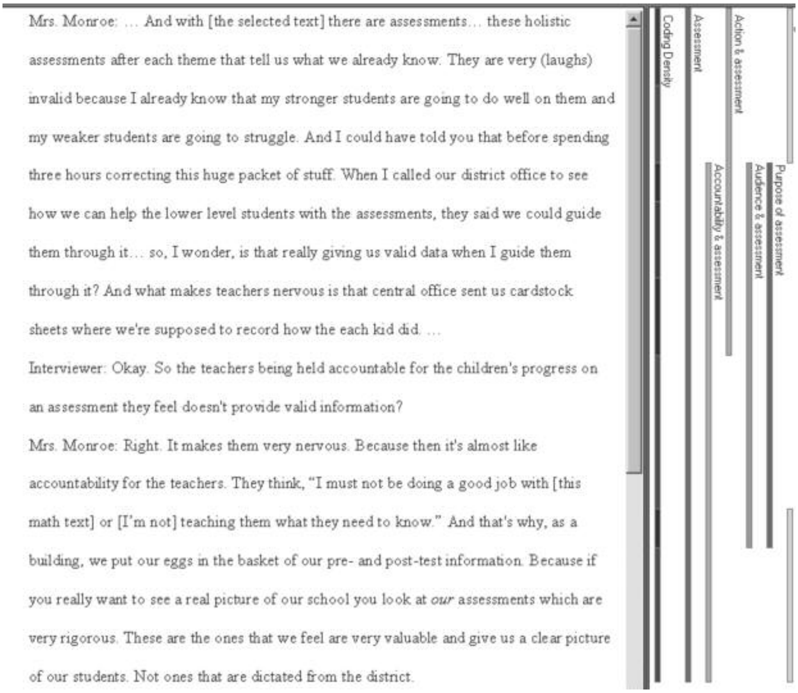
Interviewer: Okay. So the teachers being held accountable for the children's progress on an assessment they feel doesn't provide valid information?

Mrs. Monroe: Right. It makes them very nervous. Because then it's almost like accountability for the teachers. They think, "I must not be doing a good job with [this math text] or [I'm not] teaching them what they need to know." And that's why, as a building, we put our eggs in the basket of our pre- and post-test information. Because if you really want to see a real picture of our school you look at *our* assessments which are very rigorous. These are the ones that we feel are very valuable and give us a clear picture of our students. Not ones that are dictated from the district.

participants. Across three years of fieldwork, we found that assessment was a particularly powerful element in educational practice in these sites, and its connection to accountability was inextricable. We identified four thematic threads and reread the data with these themes in mind across all nine schools.

To illustrate, the interview selection next was initially coded[7] under concepts that structured the research design (Figure 1). After discussion of assessment in the case study schools, we coded the data a second time to represent new themes—audience, action, alignment, and accountability (Figure 2). Our application of codes at two stages is presented next. We constructed case studies of assessment practices in three sites, balancing attention to the themes with the site-specific issues identified by both fieldworkers in those locations and peer readers on the research team who worked to audit the assertions presented. We also used Laurel Richardson's (2000) concept of writing as inquiry.

Figure 2. Second-pass analysis in NVivo (with new codes)



Mrs. Monroe: … And with [the selected text] there are assessments… these holistic assessments after each theme that tell us what we already know. They are very (laughs) invalid because I already know that my stronger students are going to do well on them and my weaker students are going to struggle. And I could have told you that before spending three hours correcting this huge packet of stuff. When I called our district office to see how we can help the lower level students with the assessments, they said we could guide them through it… so, I wonder, is that really giving us valid data when I guide them through it? And what makes teachers nervous is that central office sent us cardstock sheets where we're supposed to record how the each kid did. …

Interviewer: Okay. So the teachers being held accountable for the children's progress on an assessment they feel doesn't provide valid information?

Mrs. Monroe: Right. It makes them very nervous. Because then it's almost like accountability for the teachers. They think, "I must not be doing a good job with [this math text] or [I'm not] teaching them what they need to know." And that's why, as a building, we put our eggs in the basket of our pre- and post-test information. Because if you really want to see a real picture of our school you look at *our* assessments which are very rigorous. These are the ones that we feel are very valuable and give us a clear picture of our students. Not ones that are dictated from the district.

This type of analysis provides the appropriate foundation for transferability (Guba & Lincoln, 1989) from the specifics of local practice to other sites and experiences because it is richly descriptive and comparative. The case studies provide rich information on how standards-based accountability and assessment practices function in context.

CONTEXT FOR INQUIRY

SAGE is a multi-faceted reform comprising four implementation pillars. In addition to the class size component, the SAGE legislation requires schools to (a) provide rigorous curricula, (b) strengthen the links between home and school through keeping the school building open for extended hours and connecting families with community resources, and (c) enhance teacher professional development and evaluation. This multidimensional approach represents the recognition that improving student achievement is a complex challenge that requires equally complex interventions, particularly in communities troubled by racism and poverty (Molnar & Zmrazek, 1994). Although initially developed to address concerns about urban poverty, SAGE is open to all Wisconsin schools, which receive $2,250 per low-income child in Grades K–3 to offset costs of implementation. Part of the SAGE legislation provides annual funding for program evaluation. We have been part of an evaluation team since 2004. This article focuses on data generated during 2006–2007.

Mirroring the theory of action suggested by many who advocate class size reduction, in the first two years of our work with these nine schools, participants reported that assessment was easier in SAGE classrooms because teacher time and attention were spread among fewer students, or, in the case of team-taught classes, the teacher shared the assessment work with a colleague. Additionally, teachers felt that the smaller groups allowed for more effective diagnosis and intervention. According to Mrs. Carter, a third-grade teacher at Bethany,

> You're able to see where the needs are and then you can work on those with individual children. . . . You're able to really individualize so much better. And when you're reading, or anything, children have so many more turns so you know right away who doesn't know their multiplication facts, whereas in the larger group you have no clue. So you can't really remediate immediately or call home or say, "There's this need."

With fewer students to work with, many elements of teaching, including assessment, became more manageable. This makes the SAGE classroom an ideal context in which to study assessment.

RESULTS

GOLDILOCKS' APPROACH TO ACCOUNTABILITY

In the fairytale of the three bears (Opie & Opie, 1974), Goldilocks moves through the bears' house, evaluating chairs, porridge, and beds in terms of too much, too little, and just right. What is good for Papa Bear or Mama Bear is often not good for Goldilocks. This story reminds us of Lilian Katz's (2000) principle of optimum effects—good for children is good only in the appropriate proportions. This can be said for accountability. Practices in schools cannot be considered in isolation. Individuals within schools engage in activities for particular reasons. Whether these reasons are purposeful and systemic, or haphazard and reactionary, they have consequences for assessment practices because they do not stand alone. Each assessment is given for a particular reason—to provide information to particular individuals and institutions. In fact, a critical aspect of validity is the degree to which an assessment provides information for its intended purpose (Moss et al., 2006).

To represent a practice-oriented perspective, we organize our analysis around the following question: What are the relationships among accountability, assessment, and practice?

In our fieldwork, we witnessed many ways in which accountability systems can benefit children and teachers. Conversely, we also witnessed many ways in which assessments and accountability systems could be improved. In the section that follows, we provide examples of assessment practices. Then, we focus on an example of what we see as best practice at Earhart Elementary School and how the administration, teachers, and students worked to design its accountability system to systematically meet the needs of its students.

"THIS PORRIDGE IS TOO COLD": ACCOUNTABILITY AND AGGREGATE ACHIEVEMENT

In all the schools we studied, staff members talked about how standardized tests influenced teachers and teaching at all levels. For example, Diane Caster, a first-grade teacher at Montford Elementary, reported,

> Everything we do [in our school] is driven by the frameworks of the testing even though our kids aren't tested until third grade. We're supposed to be laying the foundation, the basics for them. And if they don't get in first grade then they are just piling more on in second grade and if they don't have it then, by the time they get to third grade and they take the test and they don't do well, then you're judged— your school is judged. So our school is looking at the data but you're looking at the district and community. They judge the schools by how well you do on the tests, no matter how much you emphasize that that's a one-day thing. You're still being judged that way. We're being held accountable for what we do.

This teacher's perception of standardized tests and how they affect her school is not unique. Although standardized tests represent only a tiny sample of student performance for students in Grades 3–8, the ramifications of student performance on the test are profound. Though the process is informal, Ms. Caster felt that schools were judged by test scores; teachers and students at all grade levels experienced reactions to the judgments. Ms. Caster saw the tests as speaking to the district and community rather than her own practice, yet she experienced pressures exerted by the test.

This pressure was due at least in part to the expectation (set by NCLB) that schools are to improve their performance each year. This translates into what is called adequate yearly progress (AYP). According to the U.S. Department of Education (2008),

> AYP is an individual state's measure of progress toward the goal of 100 percent of students achieving to state academic standards in at least reading/language arts and math. It sets the minimum level of proficiency that the state, its school districts, and schools must achieve each year on annual tests and related academic indicators. Parents whose children are attending Title I (low-income) schools that do not make AYP over a period of years are given options to transfer their child to another school or obtain free tutoring (supplemental educational services). (p. 1)

The sanctions attached to failing to meet AYP were highly salient to administrators. According to Bill Post, the principal of small semiurban McMahon Elementary, the school eluded the dreaded list of "schools in need of improvement" because they were too small to have reliable measures. However, they still experienced local public scrutiny for their performance. It was particularly painful because they had scored well the year before:

> Everybody did pretty bad in math. AYP—47.5% is the cut off. We were at 47. So both our reading and math were .5 under, but we're exempt from being put on the "Schools In Need of Improvement" list because the number of our kids is so small, under 50 . . . . [By] race, White kids did, I think 79%, Black kids were . . . like 25%. So there's a big gap in math based on race. And [the gap between students of different socioeconomic statuses] was a big one too. The sample size is so small—it's only based on 36 kids, 47 kids took the test, only 36 were [here for the] full academic year, but to come off that 81% the previous year. Unfortunately that's how you get judged and that's what hits the papers, and then you try and explain it to people and it makes it sound like you are making excuses.

Like Ms. Caster, Dr. Post understood the audience for the assessment results to be those who read newspapers. Though this principal clearly knew the results of the standardized tests, disaggregated by race and socioeconomic status, he did not connect these results to the potential for action on his part.

Notably, both Dr. Post and Ms. Caster talked about standardized test in terms of aggregate patterns, with Dr. Post focusing on findings by race and class, and Ms. Caster drawing portraits of schoolwide achievement. Missing is attention to individual students and their needs, primarily because the state tests are not designed for instructional decision-making.

"THIS PORRIDGE IS TOO HOT": ACCOUNTABILITY AS THE BASIS FOR ACTION, PART I

Although accountability assessments are used to inform the public or the district, this is only one way to view them. Some schools use the data to inform decision-making and actions in the coming school year. Two of the best examples of this (other than Earhart, which is the focus of in-depth discussion in the next section) were Calloway and Montford, though the differences in their approaches were as significant as their similarities. In the 2005–2006 school year, student performance on the state accountability test at Calloway Elementary took a perceptible dip. In response, the principal, Mrs. Collier, spearheaded a plan for math pre- and posttests at each grade level. She also extended math instruction to

2 hours per classroom per day. She required teachers to provide extra help to one or two poorly performing students during one of their semiweekly music class periods. When third-grade writing scores on the district assessment were not up to par, Mrs. Collier stepped in and required each teacher to submit six writing samples for each student throughout the year, which Mrs. Collier personally evaluated.

In an attempt to preempt future problems, Mrs. Collier and her learning team worked to align the curriculum with district-level standards (benchmarks). At the district level, staff worked toward alignment by fostering formal partnerships with local universities. They designed classroom assessments that aligned with these benchmarks, as well as guides for "analyzing and learning from student work."

"THIS PORRIDGE IS JUST RIGHT": ACCOUNTABILITY AS THE BASIS FOR ACTION, PART II

In contrast to Mrs. Collier's top-down approach, Mary Durst, principal of rural Montford, saw herself as a "fine-tuner" of assessment practices who ensures that "it's not just something that fills up paper . . . [and] we're actually using that information to change what we are doing with kids." With increasing pressure to document student progress, teachers at Montford juggled intense demands. Mrs. Durst described how and why she and the teachers chose to work on alignment and consistency:

> Last year we found out that kindergarten wasn't doing math Trailblazers [the district mathematics program]. First grade, second grade were. But throughout the district there were some inconsistencies. Kind of like a smorgasbord. You could choose what you wanted to teach when you wanted to teach it. And especially with the impact of the standardized testing and all the frameworks we got from DPI [Department of Public Instruction], we needed some consistency. So Mrs. Felton and Mrs. Monroe both worked at the district level in creating a guide for K through 2 in monthly goals of what they should be teaching in math. So these are the "have-to's," these are the "can-do's" but they *have* to do the "have-to's." Hopefully this year the same group's going to meet together creating some common assessments which is another step the district will be taking.

Unlike Mrs. Collier, Mrs. Durst did not make a unilateral decision regarding the best course of action for teachers or students. At one of Montford's shared leadership meetings, a representative from the kindergarten team worried whether the benchmark testing at the beginning of the school year would be better timed in November, when students were more familiar with school routines. Mrs. Durst referred this issue to the school's action team, which decided that the *when* of testing is just as important as the *why* of testing. Testing was shifted to November to respond to the children's needs.

Although both administrators were proactive catalysts of increasing student achievement, the interesting distinction between these two orientations was the role of other staff members in this complex process. Although Mrs. Collier used high-stakes accountability to guide her administrative choices, she did so without the support of teachers. Her approach alienated teachers, making her strategies less effective than they could have been with the support of her staff. Many teachers felt voiceless at Calloway, unable to shape decision-making or to make their needs known. This is how third-grade teacher Marsha Delton described it:

> I like the small school, but I think it's not being run in a way that gives people a voice. Because I came from a huge school, lots of staff, lots of kids, and when I came to a small school I thought "I'll have a voice." And I think only a few people have a voice. Even on the learning team, I didn't have a voice. I don't care what anybody says, I didn't have a voice. So I'm kind of disenchanted with the building in general.

Disenchantment on the part of a teacher is a disadvantage when working to create a systemic and effective accountability system.

At Montford, there was a different conceptualization of what it meant for teachers and administrators to be accountable. Though they were still eligible for the same sanctions as any other school, this was not at the forefront of their planning efforts. Instead, they focused on constant improvement shaped by thoughtful consideration of the learning needs of all children, the goals of families, and the professional work of teachers. At Montford, the staff and administration were accountable to the community, as represented by the students, parents, and teachers, not an abstract judgmental "public."

The average CLASS score was 5.45 at Montford and 4.49 at Calloway (on a scale of 1 to 7). Interestingly, many of the dimensions measured by CLASS, including teacher sensitivity, regard for student perspectives, concept development, and quality of feedback, rely on teachers' knowledge of students' ideas, and emotional and instructional needs. Assessments, particularly those designed to inform teachers, contribute to teachers' capacity for high scores on this scale. High ratings on the quality of feedback dimension requires teachers to use informal assessment as they are teaching to identify students who are struggling, be specific about feedback given in relation to answer correctness, and engage in conversational loops for concept development. The school's average CLASS score supports our observations that at Calloway, assessments were not well integrated into the school culture. In contrast, at Montford, the staff worked collaboratively working toward making their assessments as useful and informative as possible. Moreover, at Earhart, where the average CLASS score was 5.96, there were systems in place to help teachers use assessments to improve their practice.

In many ways, this way of conceptualizing accountability had the same imagined ends: a focus on improving student performance and achievement. However, in much the same way that the *when* of testing is as important as the *how,* the audience (*who*) for accountability may be just as important as the fact that one is accountable. In the best scenarios in our fieldwork, accountability focused attention on student performance, prompting school personnel to examine outcomes of their efforts. But they also focused on building teacher capacity, recognizing that student proficiency comes from interaction with well-prepared and engaged professionals. In less desirable situations, it could be overly focused on test results without attention to productive teaching. Linking the results of assessments to action is highly dependent on the audience for assessments and the stakeholders who are held responsible for improving outcomes.

"AND SHE ATE IT ALL UP": EARHART'S RECIPE FOR JUST RIGHT

Earhart Elementary is a small diverse K–5 school nestled in the middle of a working-class neighborhood. More than two thirds of the students are classified as poor, one third are English language learners (split between Latino and Hmong), one third are African American, and one third are White. The school recently escaped threatened closure, a move that invigorated community support and teacher commitment. Four Earhart staff participated in our study in 2006–2007. Paula Walworth was in her third year as principal at Earhart. Molly Masters taught 16 kindergartners in a small but exquisitely designed space shared with another kindergarten; Tammy Helman worked with 12 first graders; and Lauren Rich taught 15 second and third graders and teamed sporadically with Betty Miller, whose classroom was next door.

Two distinct but complementary threads characterized Earhart's assessment practices. The first was a district-designed curriculum and assessment alignment that promoted coherence of district instructional practices and the state and federal accountability system. The second was the school's focus on a professional learning community that created a shared sense of purpose and responsibility in the school. It was through the joint action of these top-down and bottom-up forces that Earhart developed its specific focus on the assessment and instruction. In this section, we describe the relation between these two and how they shape the instructional practices at Earhart.

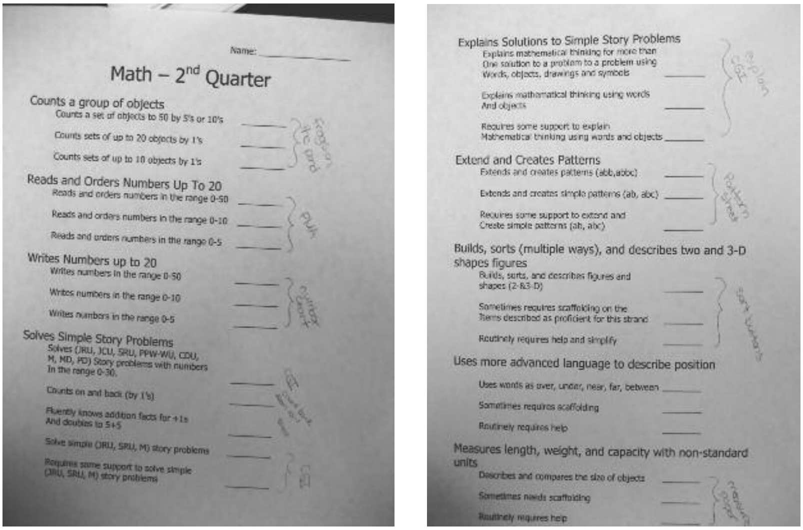
Accountability and alignment. Like many districts around the nation, the Maxwell School District had moved toward a standards-based approach to education, with more standardization of curriculum, expectations, reporting forms, and assessments. This alignment of all aspects of practice was clearly evident in educational practice at Earhart. The district curriculum standards were fully aligned with Wisconsin Academic Standards. Two district-designed, individually administered assessments, the K–3 Literacy Assessment (K3LA) and the K-2 Math Tasks (K2MT), were used to assess student learning related to the standards. Report cards had been aligned to the standards and were linked to the K3LA and K2MT, making scoring and reporting coherent. At the time of the research, district professional development had just completed a strong focus on literacy and moved onto mathematics.

Talk and action at Earhart revealed that standards, instruction, assessment, and reporting were critical and related elements of this education system. School staff regularly talked about linking their activities in the classroom with assessments and reporting forms.

Earhart's principal used district-generated data analysis to inform decision-making at the school, sharing the information with her staff. This centralized data analysis was possible because the district assessments and the report card were all computer-based, generating a district and school data bank. This resource provided perspectives on student data that Mrs. Walworth probably would not have, or could not have, completed on her own.

Ms. Masters, an incredibly organized and assessment-focused kindergarten teacher, used every spare second in the classroom for assessment. She coordinated her assessment knowledge with her practice daily by reviewing student work and documenting progress. She set up centers that mirrored the requirements of the standards-based report card. She was the paragon of alignment, working to connect all aspects of her practice. She designed the following sheets (Figure 3) to make visible connections between tasks on the K2MT and report card items :

Figure 3. Teacher-designed sheets to make visible connections between tasks on the K2MT and report card items



Name: ______

Math – 2nd Quarter

Counts a group of objects
Counts a set of objects to 50 by 5's or 10's ______
Counts sets of up to 20 objects by 1's ______
Counts sets of up to 10 objects by 1's ______

Reads and Orders Numbers Up To 20
Reads and orders numbers in the range 0-50 ______
Reads and orders numbers in the range 0-10 ______
Reads and orders numbers in the range 0-5 ______

Writes Numbers up to 20
Writes numbers in the range 0-50 ______
Writes numbers in the range 0-10 ______
Writes numbers in the range 0-5 ______

Solves Simple Story Problems
Solves (JRU, JCU, SRU, PPW-WU, CDU, M, MD, PD) Story problems with numbers In the range 0-30. ______
Counts on and back (by 1's) ______
Fluently knows addition facts for +1s And doubles to 5+5 ______
Solve simple (JRU, SRU, M) story problems ______
Requires some support to solve simple (JRU, SRU, M) story problems ______

number chart

CGI

CGI

Explains Solutions to Simple Story Problems
Explains mathematical thinking for more than One solution to a problem to a problem using Words, objects, drawings and symbols ______
Explains mathematical thinking using words And objects ______
Requires some support to explain Mathematical thinking using words and objects ______

Extend and Creates Patterns
Extends and creates patterns (abb,abbc) ______
Extends and creates simple patterns (ab, abc) ______
Requires some support to extend and Create simple patterns (ab, abc) ______

Builds, sorts (multiple ways), and describes two and 3-D shapes figures
Builds, sorts, and describes figures and shapes (2-&3-D) ______
Sometimes requires scaffolding on the Items described as proficient for this strand ______
Routinely requires help and simplify ______

Uses more advanced language to describe position
Uses words as over, under, near, far, between ______
Sometimes requires scaffolding ______
Routinely requires help ______

Measures length, weight, and capacity with non-standard units
Describes and compares the size of objects ______
Sometimes needs scaffolding ______
Routinely requires help ______

explain CGI

The teachers found the K3LA particularly well suited to the kind of daily instructional decisions they needed to make. Designed around familiar tasks like running records, these assessments nourished their instruction and helped them know more about their students. The K2MT were less informative for formative assessment because they were given only twice a year. Students were tested to the grade-level standards at the beginning of the year, which teachers found highly frustrating for many students. Echoing Ms. Caster's perception that assessments are often for purposes other than teaching, Mrs. Rich told us that they were used "so that no child is left behind." Ms. Helman saw the K2MT as what Shepard (2005) would call *benchmark assessments*, focused more on program evaluation and less on instruction: "I don't think the district would say this. It's part of measuring our progress and the quality of our teaching and how our students are achieving and learning and whether we're closing the achievement gap. It's probably to inform the public how the school system is doing."

Even in a context of close alignment, the official assessment results can have a jarring effect on teachers' psyches. Ms. Helman told us that she needed to prepare herself emotionally for student performance even though she did close assessment with her students on a daily basis. The tasks on the K3LA contrasted strongly with the books she used in instruction, "so the results can be horrifyingly different [laughs] sometimes . . .

sometimes this is a real wakeup call. I guess I feel like they need to be able to be successful on the assessments like the K3LA and the K2MT." When asked what she did when the results of the assessment diverged from what she thought a child could do, she simply said that she changed her practice. Rather than relying on her assessment of learning levels, which is developed in scaffolded instructional settings, Ms. Helman needed to look at what students could do without teacher support:

> When Destiny couldn't pass the [text reading level] 7, I thought, OK that doesn't feel like a reflection of the child I normally see. But this is what she did, so I'm going to have to take action based on it. I'm going to have to adjust my teaching. I'm probably going to do with her what I did with Matthew. For the last three weeks or so, I've met with him a second time, and I gave him extra books and just really worked on using the reading strategies. I feel like I can see that Destiny isn't doing that independently without my prompting, and she has to be able to do it by herself, so I need to get her there.

Whereas some might have discounted the results of the assessment, Ms. Helman used the divergence between her teaching and assessment experiences to prompt more intensive instruction for Destiny.

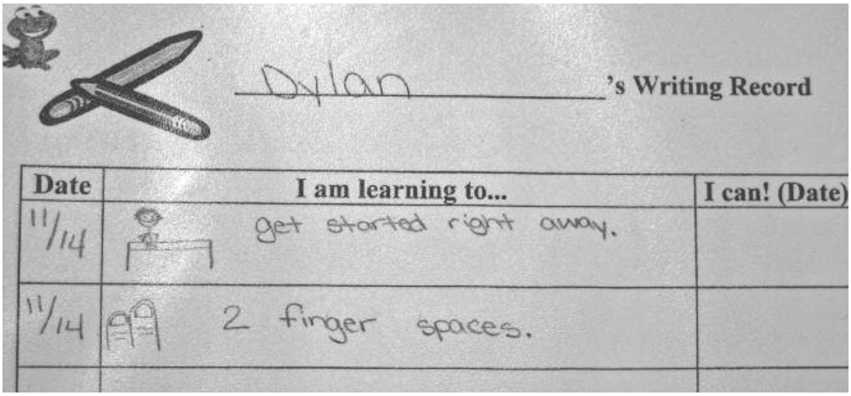
The curricular alignment undertaken by the Maxwell School District and enacted by the staff of Earhart Elementary illustrated some of the key concepts of standards-based reform. Heightened attention to the goals of instruction and systemic linkages among assessments and reporting forms set the stage for accountability. It was the second element of Earhart's unique practice that gave the power to drive accountability home. In the next section, we discuss the role that professional collaboration played in this context.

Professional learning communities and collaboration. In her third year as Earhart's principal, Paula Walworth was leading her staff in a comprehensive school reform process, professional learning communities (PLCs; DuFour & Eaker, 1998), focused on shared leadership and professional collaboration. PLCs are centered on three essential questions: (a) What do we want students to learn? (b) How will we know if they learn it? and (c) What will we do if they don't? These questions and their action-oriented premise shaped much of the work at Earhart. The staff developed shared expectations for both academic and behavioral learning and communicated them clearly to students, staff, and families. These shared expectations were communicated in common language and rules, common standards for learning and behavior, and multilingual communication. To facilitate collaborative practice, Mrs. Walworth designed schedules so that grade-level teams had weekly shared planning time, and these groups met periodically with their respective instructional resource staff.

One of the examples of shared vision related to student learning and assessment was the use of an *assessment wall*, where staff posted student data across time to understand progress. This visual depiction of learning helped staff see learning in a concrete way—a visual that moved student learning from within one teacher's head to shared documentation that often brought up questions about practice. Mrs. Walworth described how she paired use of an assessment wall with examination of instructional time between classroom and Title 1 reading teachers. Using these two together, Title teachers noticed that some high-need students were receiving reading instruction only through Title 1 pullout services, rather than having Title services supplement ongoing classroom instruction. Upon identifying this problem, classroom teachers changed their scheduling and practice to ensure that all students received classroom instruction so that Title services were in addition to regularly scheduled teaching.

Shared purpose and common tools among staff were mirrored in practices between teachers and students. Assessments were mindfully linked to instruction in the classroom to provide a strong community sense of the goals, grounding instruction in a sense-making that has purpose. Student self- and peer assessments were regularly used at Earhart as a way to communicate the goals teachers were addressing. In Ms. Masters's kindergarten, students had *I am learning to. . .* sheets in their writing binders that listed skills that they were to practice in their writing, like using two finger-spaces between words (Figure 4).

Figure 4. Ms. Masters's I am learning to . . . sheets, used as reminders for kindergarteners to self-assess, practice, and improve in their writing



Dylan's Writing Record

Date	I am learning to...	I can! (Date)
11/14	get started right away.	
11/14	2 finger spaces.	

Kindergartners were charged with assessing their success in achieving the goal and were asked to provide evidence of their accomplishments during writing conferences with Ms. Masters.

In Ms. Helman's class, students provided feedback on writing, with their teacher modeling the connections between the strategies they were learning and student use. Rich conversations around peer evaluation were made possible, in part, by the small group size of 12 first graders and Ms. Helman's modeling of supportive evaluation:

After Kiera read her book about music class, she received the following comments:

> Alison: I like how you did your pictures and how you were about music and I noticed you went "boom and boom and boom."
>
> Ms. H: What craft is that called? Using sound words!
>
> Wendy: She said, "I like music very, very much."
>
> Antoine: I like how you took your time and you didn't scribble scrabble.
>
> Ms. H: What craft did she use that you can see with your eyes? You can see it by looking. Hold up your book and show them.
>
> Kiera holds up her book and someone replies, "Bold print."
>
> Ms. H: Were you emphasizing your words? So you used bold print!!

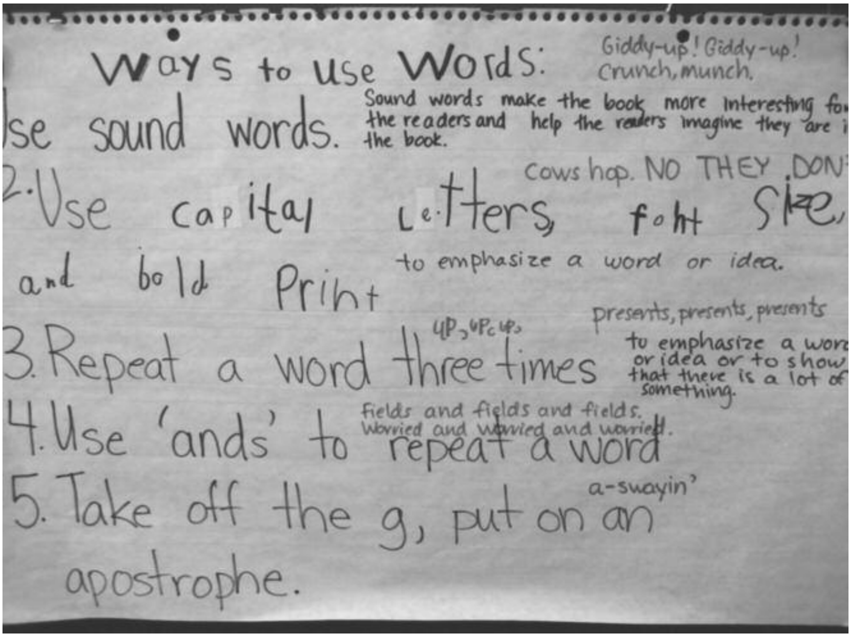
This public sharing of criteria for performance and evaluation was done in a positive collegial manner. It referenced instructional materials generated by the class that described writing crafts, seamlessly connecting instruction and assessment (see Figure 5). Ms. Helman made concrete the general comments by students, moving them beyond "I like" to specific feedback:

Self-assessment in these classrooms was a practice-based example of Black and Wiliam's (1998) assertion that:

> Self-assessment by pupils, far from being a luxury, is in fact an essential component of formative assessment. When anyone is trying to learn, feedback about the effort has three elements: recognition of the desired goal, evidence about the present position, and some understanding of a way to close the gap between the two. All three must be understood to some degree by anyone before he or she can take action to improve learning. (pp. 54–55)

Assessment was implemented at Earhart through collaborative, systemic practice. Staff worked together through consultative relationships

Figure 5. Poster made by the class that described writing crafts, seamlessly connecting instruction and assessment



Ways to use Words:

Giddy-up! Giddy-up!
Crunch, munch.

se sound words. Sound words make the book more interesting fo the readers and help the readers imagine they are i the book.

Cows hop. NO THEY DON

2. Use capital Letters, font size, and bold Print to emphasize a word or idea.

up, up, up

presents, presents, presents

3. Repeat a word three times to emphasize a word or idea or to show that there is a lot of something.

fields and fields and fields.
Worried and worried and worried.

4. Use 'ands' to repeat a word

a-swayin'

5. Take off the g, put on an apostrophe.

to maximize the resources for teaching and learning. The district's work to align standards, curriculum, assessment, and reporting provided a system that gave teacher work a more coherent quality as compared with other schools we studied and relied on teacher professional knowledge scaffolded by appropriate tools. Strong knowledge of content was generated through staff development and connected assessment practices that informed instruction. Curricular alignment, though built at the district level, had teacher-level buy-in through teacher representatives on committees that designed each element. A focus on building teacher capacity rather than teacher-proofing programming made the tools more relevant to practice.

The weak link could be seen as the report card, which appeared to serve a benchmark evaluation purpose rather than as a tool to communicate with families. The impenetrable standards-based language was difficult for even the most educated parent to understand and was exacerbated by a push at the district level that staff should limit open-ended narrative comments on the report card. This type of communication was seen as an indicator that parent-teacher conferences were needed, something outside the realm of a report card. As a tool of accountability, the report card was reframed away from being a tool for communicating with parents, and toward being a documentation instrument for school and district purposes. In working to communicate to a district audience, the report card lost its community focus, an indication of the fragility of aligned systems and the importance of balancing the needs of multiple audiences.

DISCUSSION

> Reconceptualizing assessment without at the same time reconceptualizing instruction will have little benefit … . The goal of instruction is, of course, learning, and meaningful learning is broader and deeper than the type of learning associated with most contemporary testing systems, particularly those created in response to current external accountability mandates. (Pullin, 2008, p. 335)

We began the process that led to this article as we worked to understand how class size reduction produces resources to improve instruction. We found that the current political and educational context oriented our focus to a unique intersection of issues: implementation of class size reduction, instructional practice, assessment, and accountability. Through this line of analysis, we have identified three aspects of assessment practices that affect this intersection: alignment, audience, and action.

ALIGNMENT

Systematic alignment was perhaps the most striking aspect of constructive assessment practices. Teacher participation in the alignment process was crucial for professional buy-in and made it more likely that instruction connected to assessments. Everything was more difficult in schools that lacked alignment—the system was more chaotic, assessments were seen as more of a burden. If standards, curriculum, instruction, assessment, and reporting tools were all aligned, the focus was never pulled away from the task at hand.

The degree to which this can be accomplished without diminishing teachers' professional autonomy (i.e., collaboration and shared goals instead of adopting scripted curricula) is likely to influence the ways in which the move toward alignment is received by the staff. This can be understood by thinking about how this choice affects accountability. If teachers are responsible for working together to help students learn and achieve, they are accountable to one another and to their students. If they are responsible for staying true to the text of a scripted curriculum program (i.e., reading lesson verbatim from teachers' manuals), they are accountable only to the one who has made this their job. Over the past three years of our research, many teachers whose schools have scripted curricula have reported in confidence that they often strayed from the script and supplemented it with other sources to meet the needs of their students. In some circumstances, this was frowned on by the administration. An aligned system that works with teachers, students, administrators, and families to develop goals and evaluate outcomes is preferable because the more people are involved, the more possibility there is for buy-in from all parties.

AUDIENCE

In general, the further the audience for an assessment metaphorically sits from the classroom, the less useful the assessment is for informing classroom action. The inverse is also true. It would be unwise to base state or district policy on the results of a weekly spelling test. At the heart of the alignment process was the issue of audience. Assessments served varied audiences for different purposes. Sometimes assessments were designed to inform classroom decision-making, sometimes they were used to track school efficacy. Many of these assessments were used to inform multiple audiences, including, but not limited to, state and district administrators, teachers, parents, community members, and the students themselves. When assessments were required for outside audiences and teachers could not see the relationship to their own instruction, their practice felt unaligned—the assessments seemed to intrude on precious instructional time and autonomy.

Teachers at Gallows described a bewildering number of assessments needed to satisfy district and federal audiences for the school's grant-sponsored programs. Although some were useful in their instruction, many were solely to inform the evaluation, and teachers balked at using valuable classroom time for assessment that did not inform their teaching. For assessment to be seen as a part of instructional practice, something that was not an added burden in a very busy schedule, teachers needed to see the practical connection to their work. This was the case at Earhart, where the district had developed assessments in literacy and mathematics that provided both district-level evaluation information and classroom-level instructional information.

Although serving multiple audiences was not always possible (e.g., the report cards at Earhart), when assessments were administered, it had to be clear who would benefit from their results. Although using the term *benefit* in relation to assessment may seem out of place, if the purpose of assessment is to improve, then the desired outcome—the benefit of assessments—must be explicit. Moreover, if students are to benefit from taking a particular assessment, the audiences of the assessments must be held accountable for ensuring that students receive that benefit.

ACTION

The final aspect is action, or the degree to which educators felt they could act on assessment information. The practices at Earhart promoted action that linked assessments, instruction, collaboration, and professional development. The systemic nature of the district's approach facilitated this and was taken up by a school staff hungry to take an active role in planning.

Action was not always positive. The pressures teachers felt to have children perform on assessments sometimes pushed them to resort to bribery. Mrs. Manchester, a kindergarten teacher at Wellstone Blvd., needed children to count to 100 by the end of the school year. An initial assessment found only three students who met this goal. She made it more likely by announcing to her students,

> Last week I tested people who were ready to count to 100, and Miguel, Thuyet, and Jared counted to 100. And those three boys are going to get a candy bar at the end of the day. [Kids gasp and look around the group.] I went out and bought candy bars for people who counted all the way to 100. And tonight you're all going to get a sheet to practice your counting, and when you're ready to take the test, if you count all the way to 100, I'm going to give you a candy bar!

Focused attention on assessed need can jumpstart student learning, particularly when it is followed by targeted instruction (Black & Wiliam, 1998). In this case, using candy as an incentive and outsourcing the learning to home certainly missed some in-class learning opportunities that could have enriched education in this class. The use of rewards is unlikely to increase the learning in this classroom, in fact will likely reduce motivation (Kohn, 1999; Lepper, Green, & Nisbett, 1973).

CONCLUSION: ACCOUNTABILITY THAT'S "JUST RIGHT"

In schools where accountability frameworks were focused on state-level standardized assessments, how the staff perceived these tests influenced the degree to which the results could be used to benefit students. Evasive or defensive reactions were not constructive in improving student achievement. Dr. Post and Mrs. Caster were so concerned with assessment reporting that they had little energy to think about how instructional practice might be reconceptualized.

In schools where approaches to accountability were based on collaboration and constructive action to improve student achievement and ensure student learning, like Earhart, assessment, instruction, and accountability flowed together recursively. Everyone was responsible for student learning, and how to best allocate resources to positively affect student learning was a community-wide concern.

Coherent and collaborative assessment practices were more likely to take place in schools with higher achievement. In these contexts, there were explicit connections through assessments to varied communities of interest: district, school, teachers, students, and families. These connections increased the likelihood that the assessments met the needs of these audiences and that they prompted some kind of action. Notably, these connections also contributed to higher ratings on CLASS. Action was the lynchpin of high-quality assessment. In supportive assessment systems, teachers had tools that they understood and that they could use to improve practice. This improvement was a contingent one, related to the needs of their students this year. In contrast, assessment in lower quality classrooms took place in disjointed systems that focused primarily on summative rather than formative assessment. In these schools, teachers had tools to find out where students were, but this knowledge was not connected to instructional action. Goldilocks would describe troubled assessment and accountability systems as “too big” or “too little,” and productive examples as “just right.”

The promises of standards and accountability have captured national attention and many education resources. Schools have labored to incorporate them into their practice, some successfully, others less so. In an interesting example of synchronicity, as we were finishing this article, a special committee of the National Academy of Education published a policy brief calling for a reinvention of the accountability project, with stronger links to student learning rather than test scores:

> Accountability systems have yet to bring about the hoped for improvements in learning. We urge state and federal leaders to fully review the status and effects of test-based accountability policies. The intention of these policies—to focus attention on student learning, make schools more responsible, and provide

> guidance for educational improvement—are in the best interest of the country. (National Academy of Education [NAE], 2008)

The NAE white paper on accountability accentuates many of the themes we identified in this article—that a focus on accountability without attention to instructional and assessment resources quality is inherently flawed. We agree with their conclusions and hope that there will be continued assessment and accountability for the accountability systems currently in place in the United States.

Notes

1. The Assessment Reform Group developed from the British Educational Research Association's Policy Task Force on Assessment in 1989. Its membership has evolved, but the foundational members are Paul Black, Patricia Broadfoot, Richard Daugherty, John Gardner, Caroline Gibbs, Wynne Harlan, Mary James, Gordon Stobart, and Dylan Wiliam. For more information, go to http://www.assessment-reform-group.org/.
2. All names for participants and sites are pseudonyms.
3. While instruction went on in as typical a manner as possible, only students with signed permission forms were videotaped. When a child inadvertently came within the video screen, he or she was digitally removed.
4. Interview protocols can be found on the WCER SAGE evaluation Web site in the 2006–2007 Final Report at http://varc.wceruw.org/sage/.
5. CLASS training must be completed with a certified trainer who provides two days of video examples of CLASS domains and dimensions, as well as multiple opportunities to rate sample videos. To be certified, a rater must meet an 80% reliability with master coders.
6. For more information on CLASS, see http://www.classobservation.com/. Because it was not a focus of this analysis, we did not include a detailed discussion of this valuable tool in this article.
7. It should be noted that not every word in this excerpt is relevant to each code. As a rule, we code in larger chunks to provide sufficient context for ideas.

References

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box.* Cambridge, England: Cambridge University Faculty of Education.

Assessment Reform Group. (2002). *Assessment for learning: Ten principles.* Cambridge, England: Cambridge University Faculty of Education.

Baker, E., & Linn, R. (Eds.). (2004). *Validity issues for accountability systems.* New York: Teachers College Press.

Biddle, B. J., & Berliner, D. C. (2002). Small class size and its effects. *Educational Leadership, 59*(5), 12-23.

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

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 20–50). Chicago: National Society for the Study of Education.

Blatchford, P. (2003). *The class size debate: Is smaller better?* Maidenhead, England: Open University Press.

Brown, C. P. (2007). Unpacking standards in early childhood education. *Teachers College Record, 109*(3), 635–668.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24,* 305–331.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047–1085.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106*(6), 1145–1176.

DuFour, R., & Eaker, R. (1998). *Professional learning communities at work: Best practices for enhancing student achievement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Ellwein, M. C., & Glass, G. V. (1991). Testing for competence: Translating reform policy into practice. *Educational Policy, 5,* 64–78.

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*, 321–368.

Forster, M., & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 51–73). Chicago: National Society for the Study of Education.

Frederiksen, J. R., & White, B. Y. (2004). Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 74–104). Chicago: National Society for the Study of Education.

Gardner, D. P. (1983). *A nation at risk.* Washington, D C: National Commission on Excellence in Education, U.S. Department of Education.

Goldstein, L. S. (2008). Kindergarten teachers making "street level" education policy in the wake of No Child Left Behind. *Early Education and Development, 19*, 448–478.

Graue, M. E., & Walsh, D. J. (1998). *Study children in context: Theory, methods, and ethics.* Thousand Oaks, CA: Sage.

Grissmer, D. (1999). Class size effects: Assessing the evidence, its policy implications, and future research agenda. Conclusion. *Educational Evaluation and Policy Analysis, 21*(2), 231–248.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation.* Newbury Park, CA: Sage.

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141–166). New York: Teachers College Press.

Katz, L. G. (2000). *Last class notes.* Retrieved December 18, 2008, from http://ceep.crc.uiuc.edu/pubs/katzsym/katz.pdf

Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes.* New York: Houghton Mifflin.

Koretz, D. (2008). *Measuring up. What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report No. 551). Los Angeles: Center for the Study of Evaluation, University of California.

Lepper, M., Greene, D., & Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic rewards. *Journal of Personality and Social Psychology, 28*, 129–137.

Linn, R. L, Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing*. Charlotte, NC: Information Age Press.

Molnar, A., & Zmrazek, J. (1994). *Improving the achievement of Wisconsin's students. Urban initiative task force recommendations and action plan* (Bulletin No. 95079). Madison: Wisconsin Department of Public Instruction, Office of Policy & Budget.

Moss, P., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

National Academy of Education. (2008). *Standards, assessments and accountability: Education policy briefing sheet*. Retrieved from http://www.naeducation.org/White_Papers_Project_Standards_Assessments_and_Accountability_Briefing_Sheet.pdf

National Research Council. (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academy Press.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*, 155–175

Nichols, S., & Berliner, D. (2007). *Collateral damage. How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard University Press.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 (2002).

Odden, A., & Archibald, S. (2000). *Reallocating resources: How to boost student achievement without asking for more*. Thousand Oaks, CA: Corwin Press.

Opie, I., & Opie, P. (1974). *The classic fairy tales*. Oxford, England: Oxford University Press.

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Pianta, R., La Paro, K., & Hamre, B. (2006). *CLASS: Classroom assessment scoring system manual K–3 version*. Charlottesville, VA: Center for the Advanced Study of Teaching and Learning.

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *CLASS Classroom assessment scoring system manual K–3*. Baltimore: Paul H. Brookes.

Pullin, D. (2008). Assessment, equity, and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. Jones Young (Eds.), *Assessment, equity and opportunity to learn* (pp. 333–352). New York: Cambridge University Press.

Richardson, L. (2000). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 923–948). Thousand Oaks, CA: Sage Publications.

Rothstein, R. (2008). *Grading education: Getting accountability right*. New York: Teachers College Press & Economic Policy Institute.

Rothstein, R., Jacobsen, R., & Wilder. T. (2008). *Grading education: Getting accountability right*. Washington, DC, and New York: Economic Policy Institute and Teachers College Press.

Ryan, K. E., & Shepard, L. A. (Eds.). (2008). *The future of test-based educational accountability*. Oxford, England: Taylor Francis.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *The handbook of research on teaching* (4th ed., pp. 1066–1101). Washington, DC: American Educational Research Association.

Shepard, L. A. (2005, October). *Formative assessment: Caveat emptor*. Paper presented at the ETS invitational conference, The Future of Assessment: Shaping Teaching and Learning, New York, NY.

Shepard, L. A. (2008). A brief history of accountability testing, 1965–2007. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 25–46). New York: Routledge.

Shepard, L. A., & Dougherty, K. C. (1991, April). *Effects of high-stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago, IL.

Shepard, L. A., Hammerness, K., Darling-Hammond, L., & Rust, F. (with Snowden, J. B., Gordon, E., Gutierrez, C., & Pacheco, A). (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275–326) San Francisco: Wiley.

Shepard, L. A., Kagan, S. L., & Wurtz, E. (1998). *Principles and recommendations for early childhood assessments.* Washington, DC: National Education Goals Panel.

Smith, M. L. (2003). *Political spectacle and the fate of American schools.* New York: Routledge.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1991). *The role of testing in elementary schools* (CSE Technical Report No. 321). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–267) Bristol, PA: Falmer Press.

Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*, 324–328.

U.S. Department of Education. (2008). *Adequate yearly progress.* Retrieved November 30, 2008, from http://answers.ed.gov/

Valli, L., & Chambliss, M. (2007). Creating classroom cultures: One teacher, two lessons, and a high-stakes test. *Anthropology and Education Quarterly, 38*, 57–75.

Wills, J. S., & Sandholtz, J. H. (2009). Constrained professionalism: Dilemmas of teaching in the face of test-based accountability. *Teachers College Record, 111*(4), 1065–1114.

ELIZABETH GRAUE is a professor in the Department of Curriculum & Instruction and associate director of Faculty, Staff, and Graduate Development at the Wisconsin Center for Education Research. She is interested in issues of early childhood policy, including kindergarten practices, class size reduction, and assessment. Recent publications include: Graue, E., Rauscher, E., & Sherfinski, M. (2009, December). The synergy of class size reduction and classroom quality. *Elementary School Journal*; and Graue, E. (2009, November). Reimagining kindergarten. *School Administrator,* 10–14.

ERICA JOHNSON is a PhD candidate in the Department of Curriculum & Instruction at the University of Wisconsin–Madison. Her research interests involve religious pluralism and controversy in U.S. public schools. Recent publications include: Graue, E., Rauscher, E., & Sherfinski, M. (2009, December). The synergy of class size reduction and classroom quality. *Elementary School Journal*; and Graue, E., & Rauscher, E. (2009, May). Researcher perspectives on class size reduction. *Educational Policy Analysis Archives, 17*(9).