American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. X–X
DOI: 10.3102/0002831208316955

Improving Teachers' Assessment Practices Through Professional Development: The Case of National Board Certification

Mistilina Sato
University of Minnesota, Minneapolis
Ruth R. Chung
Linda Darling-Hammond
Stanford University, California

This study examines how mathematics and science teachers' classroom assessment practices were affected by the National Board Certification process. Using a 3-year, longitudinal, comparison group design, evidence of changes in teachers' classroom practice were measured on six dimensions of formative assessment. The National Board candidates began the study with lower mean scores than the comparison group on all six assessment dimensions; had higher mean scores on all dimensions by the second year, with statistically significant gains on four of the dimensions; and continued to demonstrate substantially higher scores in the third year. Pronounced changes were in the variety of assessments used and the way assessment information was used to support student learning. National Board candidates attributed changes in practice to the National Board standards and assessment tasks. Comparison group teachers who showed noticeable changes in practice described professional development experiences similar to those supported by the National Board Certification process.

KEYWORDS: teacher education/development, assessment, science education, mathematics education

Educators, researchers, and policy makers are increasingly interested in identifying practices that contribute to improved student learning, performance, and achievement. Classroom assessment may be a particularly productive, if generally underused, lever for transforming practice in ways that support student learning. Black and Wiliam's (1998a) review of several hundred empirical articles concerned with classroom formative assessment reported consistent learning gains for students when teachers use assessment

practices that support learning as well as surprisingly large effect sizes for well-developed formative assessment practices.

In this regard, the National Board Certification process offers a promising site for investigation, given recent evidence that it may serve both as a tool for *identifying* more effective teachers (Bond, Smith, Baker, & Hattie, 2000; Cavaluzzo, 2004; Goldhaber & Anthony, 2005; Smith, Gordon, Colby, & Wang, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004) and as a means for *developing* accomplished teaching, because of the professional learning that appears to accompany engagement in the process. In particular, a number of studies have found that teachers identify their classroom assessment strategies as undergoing change as a result of their engagement in the certification process (Athanases, 1994; Lustick & Sykes, 2006; Sato, 2000; Tracz et al., 1995; Tracz, Sienty, & Mata, 1994). By examining the critical case of the National Board Certification process as a professional learning opportunity for teachers, this study explores how this and similar professional development experiences can potentially improve everyday formative assessment practices in the classroom.

Previous interview and survey studies of teacher learning related to participation in the National Board Certification process and many other professional development experiences have relied on self-reports of change. This study explores how we might directly measure changes in teachers' practices by seeking confluence among a variety of data sources drawn directly from teachers' classrooms and experiences over a period of time.

This study tracked National Board candidates' assessment practices over 3 years—a year prior to pursuing National Board Certification, a year of candidacy, and the post-candidacy year—along with those of a comparison group of teachers who were interested in pursuing National Board Certification but who postponed their candidacy until the study was completed. Nine National Board candidates in middle and high school science and mathematics certificate areas (Early Adolescence and Adolescence and

---

Mistilina Sato is an assistant professor of teacher development and science education at the University of Minnesota, 159 Pillsbury Dr., SE, Peik Hall, Minneapolis, MN 55455; e-mail: *msato@umn.edu*. Her research focuses on teacher professional development, especially as it relates to phronesis and practical reasoning, and formative assessment integrated with instruction in science classrooms. She has coauthored books on case methods in developing teachers' mentoring practices and on teacher development related to formative assessment practices. In 2007, Sato was awarded the early career research award by the Teaching and Teacher Education Division of the American Educational Research Association.

Ruth R. Chung is a postdoctoral scholar at Stanford University, School of Education, 505 Lasuen Mall, Barnum 111, Stanford, CA 94305; e-mail: *rchung@stanford.edu*. Her professional interests include teacher quality and learning, teachers' classroom assessment practices, performance assessment, and student assessment.

Linda Darling-Hammond is the Charles E. Ducommun Professor of Education at Stanford University, School of Education, 520 Galvez Mall, 326 CERAS Building, Stanford, CA 94305; e-mail: *ldh@stanford.edu*. Her research, teaching, and policy interests focus on teaching quality, school reform, and educational equity.

Young Adulthood) and 7 similarly experienced non–National Board teachers in the same fields participated in all 3 years of the study. We derived classroom evidence about six dimensions of assessment practice from videotapes of teaching, lesson plans, and student work collected over the years of the study and supplemented these data with teacher interviews and student surveys to document the extent and nature of changes in practice.

Background of the Study

The National Board for Professional Teaching Standards (referred to throughout as the National Board) was created in 1987 with a mission to "establish high and rigorous standards for what accomplished teachers should know and be able to do, to develop and operate a voluntary national system to assess and certify teachers who meet those standards, and to advance related education reforms—all with the purpose of improving student learning" (Baratz-Snowden, 1990, p. 19). To achieve National Board Certification, candidates must complete a rigorous two-part assessment. The assessment includes a portfolio completed by the teacher at the school site, which incorporates student work samples, videotapes of classroom practice, and extensive written analyses and reflections based upon these artifacts. The portfolio is meant to allow teachers to present a picture of their practice as it is shaped by the particular needs of the students with whom the teachers work and the particular context of the teacher's school. The assessment also includes a set of exercises completed at a local assessment center during which candidates demonstrate both content knowledge and pedagogical content knowledge through tasks such as analyzing teaching situations, responding to content matter prompts, evaluating curriculum materials, or constructing lesson plans.

Since the National Board began certifying teachers, there has been interest in the question of the extent to which the certification process influences teachers' thinking, learning, and practice. Studies examining teachers' reactions to the National Board assessment process, along with testimonials from individual teachers, have consistently reported teachers' becoming more conscious of their teaching decisions and changing their practices as a result (see, e.g., Chittenden & Jones, 1997; Sato, 2000; Tracz et al., 1994; Tracz et al., 1995).

A common thread running through the research is teachers' reported change in their understanding of assessment in their classrooms. For example, Athanases (1994) reported that almost 90% of the teachers in his study indicated that their classroom practices improved as a result of their participation in the National Board portfolio assessment process. In particular, teachers felt their ability to assess student learning improved. In a longitudinal, quasi-experimental study that investigated learning outcomes for high school science teachers who pursued National Board Certification, Lustick and Sykes (2006) found that one of the most significant areas of learning was related to the teaching standard for assessment. Interview data with teachers suggested that this standard helped the teachers deepen their conception of assessment, leading them to see instructional purposes of assessment beyond summative test results and grades.

A growing body of research has found that the extent to which teachers embed formative assessment practices in their everyday classroom interactions is strongly related to student learning and is associated with improved student achievement. Black and Wiliam (1998b) define formative assessment as including two interrelated parts: first, activities undertaken by the teacher and the students as a means of collecting information about the students' understanding or progress and, second, the use of this information to modify teaching and learning activities by the teacher, the students, or both.

Within this broad domain, research on learning flags the importance of providing formative or diagnostic information to teachers and students, providing clear expectations and goals for learning, creating coherence between assessment and curriculum, and supporting metacognitive practices. For example, research on classroom-based assessment suggests that greater student learning and higher task performance are achieved by providing task-oriented feedback to students (Butler, 1987; Crooks, 1988) and by eliciting information from students through assignments and discussion as a means of gauging where students are in their progress toward a goal (Duschl & Gitomer, 1997).

Research on learning also suggests that understanding is strengthened when the learners are asked to take an active part in determining what they understand and how they came to that understanding, as well as what they still need to learn (National Research Council, 2000). Classroom practices that aid this kind of metacognition include peer- and self-assessment, reflection on one's own progress and determining what needs further improvement, and activities geared toward allowing students to make sense of new concepts through talk or writing, which allow the teacher to gather information on student understanding to guide his or her next steps (Palincsar & Brown, 1984; Scardamalia, Bereiter, & Steinbach, 1984; White & Frederiksen, 1998). Structuring these kinds of opportunities is formative assessment practice.

Finally, making the goals for student learning and performance explicit is a fundamental premise of national and local work on setting standards within disciplinary areas. At the classroom level, the extent to which the teacher sets clear learning and performance goals helps both students and teachers assess what the students have learned and where the students need continued work and support (Stiggins, 1994; Wiggins & McTighe, 1998). When taken together, these activities are not solely about assessment but represent a way of interacting with students that is purposeful toward learning and capitalizes on everyday interactions in the classroom.

A Framework for Examining Assessment Practices

As we have described, assessment is tightly embedded in many acts of teaching; thus, a multi-dimensional approach to examining it is necessary. We sought to develop an analytic framework that conceptualizes the variety of actions and decisions that go into a teacher's assessment practice, as well as the variety of roles that assessment plays in the classroom.

1. Views and uses of assessment
   In the teacher's view, what counts as assessment?
   How is assessment information used in this classroom (for student learning and for guiding instructional decisions)?
   How is assessment viewed and used by the class?
2. Range, quality, and coherence of assessment methods
   What is the variety of assessment methods the teacher uses for purposes of gathering information about student progress?
   What is the quality of the assessment methods?
   Are the assessment methods consistent with the learning goals?
   Are the assessment methods strategically used to help further student learning?
3. Clarity and appropriateness of goals and expectations for learning
   Are the learning goals and criteria of quality clearly articulated?
   Are the learning goals conceptually important based on current thinking in the field?
   Are the learning goals important and appropriate for the students (developmentally, readiness, interests)?
4. Opportunities for self-assessment
   Are there opportunities for student self-assessment (e.g., reflecting on performance, monitoring progress over time, predicting future performance, determining what needs further improvement, reflecting on one's metacognition)?
   How does the teacher scaffold or guide the student self-assessment?
   What is the overall quality of these opportunities (e.g., relationship to learning goals, time allowed, consistency in opportunities, how information is used by students and teacher)?
5. Modifications to teaching based on assessment information
   Does the teacher take into account prior knowledge of the students?
   Does the teacher demonstrate flexibility and responsiveness to the students' needs and interests during instruction?
   How is assessment information used to guide future instructional decisions?
6. Quality and appropriateness of feedback to students
   How specific is the feedback to the task or assignment?
   Does the feedback prompt students to take further action?
   Is whole-class feedback tailored to specific activities or students' needs?

Figure 1. **Six dimensions of formative assessment.**

Based on the formative assessment literature and an analysis of the expectations for classroom assessment practice outlined in the National Board standards for middle and high school science and mathematics teachers, we identified six dimensions of formative assessment (see Figure 1). Each of the six dimensions was further expanded into several indicators, which were specified on a 5-point rubric describing the features of practice and the quality of performance for each of the levels. For example, under Views and Uses of Assessment, the ratings for one indicator ranged from a 1 where *assessment information is only used for evaluating students' performance at the end of an instructional sequence (e.g., testing or grading at the end of a unit)* to a 5 where *assessment information is used for summative evaluation purposes, as an opportunity for students to improve the quality of*

their work or their understanding of the ideas or skills, and as a means to help students understand the progress of their own learning.

The rubrics were used to analyze the degree of use and the quality of formative assessment in participants' classroom practices at multiple points in time over the 3 years. This framework allowed us to assess the extent to which teachers in each of the comparison groups (a) changed their views about the purposes and uses of assessment in their classroom, (b) increased the variety and frequency of their formative assessment practices, (c) gained greater clarity about goals and expectations, (d) developed greater coherence between their curriculum and assessment goals and strategies, (e) increased the use of strategies that support student meta-cognition about their own learning, and (f) used the results of their assessments to inform and modify their teaching.

Research Design and Methods

The Sample

Potential participants, who were middle or high school mathematics or science teachers, were solicited through contacts offered by Stanford University's support group for National Board Certification candidates. Sixty teachers interested in pursuing certification and considering participation in the study attended research project orientation meetings in January 2003. The project was entitled "Examining Changes in Classroom Practices" to avoid alerting candidates to the focus on formative assessment, which could have influenced their practice.

Typically, as teachers learn more about the certification requirements and commitments, many decide not to pursue the process, given the amount of work involved and events that occur in their personal or professional lives. We also knew that the demands of the study would be substantial for the participants and thus expected attrition of teachers over the course of the 3 years. We aimed for 20 teachers assigned to each of the National Board Certification and the non–National Board groups and after attrition ended the study with 16 participants: 9 National Board candidates and 7 non–National Board participants. Some of those who withdrew from the study moved out of the area; others left the classroom to pursue a master's degree, take another job, or start a family. A number indicated that their classroom and school responsibilities were too demanding for them to remain in the National Board Certification process or to complete the data collection protocols associated with the study.

The study began with a baseline year of data collection about assessment practices for all the participating teachers. In the second year, the same data collection strategies were used while half of the teachers pursued National Board Certification and the half not pursuing certification remained in the comparison group. The decision of whether to pursue certification was left to the individual teachers; however, as an incentive for some teachers to wait, we offered financial support for their later certification fee. In guiding

Table 1
Participants' Characteristics and School Contexts

	National Board	Non–National Board
Teacher characteristics		
Average years of teaching experience	9.9	11.7
Subject area	8 science 1 mathematics	5 science 2 mathematics
Grade level	3 middle school 6 high school	4 middle school 3 high school
School characteristics		
Average Academic Performance Index	6.0	9.0
Academic Performance Index range	1 to 10	7 to 10
Average total enrollment	1,378	1,166
Average school enrollment by race/ethnicity (%)		
African American	11.0	3.1
American Indian or Alaska Native	0.5	0.5
Asian	17.1	32.4
Filipino	4.4	4.6
Hispanic/Latino	23.2	7.7
Pacific Islander	1.2	1.3
White, not Hispanic	41	49.5
Multiple or no response	1.7	1.5
Average English learner enrollment (%)	13.6	8.9
Average percentage fully credentialed teachers (%)	87.7	95.1[a]
Average pupil-to-teacher ratio	21.7	19.4
Average class size	28.9	26.5
Average participation in free and reduced-price meals (%)	23.3	12.1
Average participation in CalWORKs (formerly Aid to Families With Dependent Children) (%)	4.8	2.4

Note. Data on schools were retrieved from the California Department of Education Dataquest Service for the 2003-2004 academic year, the first year of the study. School characteristics did not change appreciably over the 3 years of the study.
[a]One private school in this group was excluded from this average because it does not require teachers to be credentialed.

teachers' decisions, an effort was made to match the groups on teaching experience and the type of school in which they taught.

As Table 1 shows, the two comparison groups were similarly experienced at the beginning of their participation in the study: National Board candidates taught for an average of 9.9 years, while non–National Board participants had an average of 11.7 years of teaching experience. Both groups had a preponderance of science teachers. Three of the National Board candidates and 4 of the comparison group participants taught at the middle school level. One of the non–National Board participants taught in a small, religious private school.

The National Board candidates taught in higher-need schools with lower achievement test scores as measured by the state Academic Performance Index rating. These schools also had higher proportions of racial/ethnic minority students, students enrolled in the free and reduced-price lunch program, and English-language learners. The National Board candidates also experienced larger class sizes and had fewer credentialed teachers in their schools. All of the above suggest that they worked in more challenging teaching conditions.

Data Collection

Data collection included (a) twice yearly videotapes of a sequence of three to five classroom lessons; (b) written responses to questions about the videotaped lessons that probed the sources of the curriculum materials, intended learning goals, planned assessment techniques, and an evaluation of the success of the lesson by the teacher; (c) collection of student work samples and lesson plans from the same unit of study as the videotaped lessons; (d) twice yearly teacher interviews about their practices and assessment approaches in general and specifically related to the unit of study described in the lesson plans and shown in the videotapes; (e) student and teacher surveys about teaching and assessment practices collected near the end of each of the 3 academic years of the study; and (f) final reflective interviews with teachers about perceived changes in practices and influencing factors.

Participants were paid a stipend for each data packet that they submitted to the project staff, ranging from $200 in the first year to $300 in the third year. Those who chose to pursue National Board Certification also received a $1,000 subsidy for their National Board Certification fees. Project staff provided occasional technical assistance with videotaping.

Analysis of Classroom Data

The analyses of classroom data were guided by a set of rubrics on the six dimensions of formative assessment described earlier, resulting in data packet scores for each candidate on each dimension. The rubric for each dimension of formative assessment included guiding questions and performance indicators organized into a 5-point scale to focus the scorer on the factors to consider in evaluating each dimension. Data packet scores were then analyzed for evidence of changes in teachers' practices across the 3 years of the study.

Scoring Process

Through extensive conversations and a series of moderated scoring sessions, the research team spent a significant amount of time refining the rubrics and developing a scoring process that could be used reliably. Benchmark data packets exemplifying each of the score levels were established. Scorers were asked first to review all of the written documentation in the data packet, starting with the Lesson Plan Overview, then the Video Cover Sheets, then the Student Work Sample Cover Sheet, the Student Work Samples, and the

interview summary. Scorers took notes about supporting evidence on the rubrics while reviewing this documentation. After reviewing all of the written documentation, scorers viewed the video. After evaluating several scoring procedures, we found that reliable ratings could be obtained based on a sample of the teacher's classroom videotape, which included the first several minutes of a lesson and subsequent segments, especially at points of transition in lesson activities.

A rubric score for each dimension was assigned to a data packet considering the evidence from all data sources in the packet (teacher descriptions, lesson plans, student work, and videotape) and the teacher interview. Evidence was weighed based on how strong and convincing it was, rather than the mere quantity of artifacts. A teacher could have one incidence of a practice and score high on the rubric if the assessor judged the performance to be strong and of high quality. Using the same logic, multiple instances of a weaker assessment practice would score lower on the rubric. We chose to base scoring judgments upon the preponderance of evidence rather than an additive model or averaging scores of discrete descriptors of practice because we thought that this approach would best capture the complexity of formative assessment practices that are integrated in overall instruction. The overall score for the rubric was assigned; then supporting evidence for the rating was noted on the rubric descriptors.

Rater Assignment and Interrater Reliability

Each teacher was assigned an identification code in order to maintain anonymity during the scoring process. Since the study sought to identify changes over time in practice, first raters were assigned to score data packets of the same teacher over time with the rationale that familiarity with a teacher's practice would allow the scorer to see qualitative differences over time that would assist us in identifying participants for in-depth case studies. Second raters were assigned randomly. While most raters were scoring data packets blind, this was not always possible, as some research project staff overlapped with staff working within a National Board support program at the university and some assisted teachers in completing their videotaping. To minimize bias in the scoring process, we assigned raters blindly when possible, and we maintained a rigorous double scoring protocol to check and maintain reliability in the scoring process.

To establish interrater reliability, 66% of the first-year data packets, 41% of the second-year data packets, and 41% of the third-year data packets were scored independently by multiple raters. Consensus estimates (see Table 2) showed that 53% of overall scores were exact matches (differing by less than 0.25 of a point), and 94% of scores differed by no more than 1 point on the 5-point rubric. Spearman-Brown reliability estimates, calculated for all data packets across all 3 years, show an overall level of reliability of 0.96, with subscore reliabilities ranging from 0.93 on Views and Uses of Assessment to 0.74 on Clarity and Appropriateness of Goals and Expectations for Learning, well within accepted limits.

Table 2
Assessor Reliability and Standard Error of Scoring of Rubric Scores (All Years)

Rubric	Correlation	Reliability[a]	Standard Error of Scoring
Views and uses of assessment	.868	.929	.222
Range, quality, and coherence of assessment methods	.750	.857	.318
Clarity and appropriateness of goals and expectations for learning	.584	.737	.351
Opportunities for self-assessment	.804	.891	.304
Modifications to teaching based on assessment information	.666	.800	.382
Quality and appropriateness of feedback to students	.761	.864	.389
Overall[b]	.778	.963	.815

[a]To estimate assessor reliability, the Pearson-Brown formula was utilized to obtain the correlation between rubric scores.
[b]Represents the assessor reliability across all sets of double scores for all rubrics, for all data packets, across all 3 years.

Analysis of Data Packet Scores

Each teacher was expected to submit a total of six data packets, two each year for 3 years. Analyzing and scoring the data packets allowed us to identify changes in teachers' conceptions of assessment and their classroom assessment practices over the 3 years of the study. In addition, the six rubrics used to score the data packets allowed us to pinpoint specific changes in teachers' practice. In order to examine data packet score trends across the two groups over the 3 years of the study, we used *t* tests to compare mean scores of each group for each year. We also conducted *t* tests of mean score differences for each group on the data packet scores between the first, second, and third years.

Because the group scores tend to mask the trajectories of change for individual teachers in both groups, we also analyzed how individual teachers showed evidence of growth or change in their conceptions of assessment and in their classroom assessment practices. Using the data packet scores across the 3 years of the study, we were able to track changes in individual teachers' formative assessment practices over time.

Analysis of Student Surveys

Student surveys were administered near the end of each academic year. Surveys were completed by students in the class that was videotaped and by

a second class the teacher identified as most comparable to the videotaped class. The student surveys asked students to report on the frequency of their participation in the class; the frequency of use of measurement tools, manipulatives, calculators, computers, or electronic probeware; and their teachers' emphasis on 33 different classroom activities.

Four free-response questions asked students to describe (a) how they know what their teacher wants them to learn in the class (goals for learning), (b) the kinds of discussions they have with their teacher about their work (oral feedback), (c) the kinds of written comments their teacher makes on their work (written feedback), and (d) whether they get to evaluate their own work or the work of other students (self- and peer-assessment). The 5,922 student free-response answers across the 3 years were analyzed by coding responses into 71 categories (22 for goals, 22 for written feedback, 25 for oral feedback, and 2 for self- and peer-assessment). Categories representing more than 5% of student responses were selected for further analysis. These categories (8 for goals, 6 for written feedback, and 6 for oral feedback) were analyzed for changes in the frequency for each teacher and for the two groups over the 3 years of the study. For the self- and peer-assessment questions, which had substantially fewer responses than the other free-response questions, teachers' practices were characterized into 3 categories.

We analyzed mean student survey ratings and mean changes in scores for assessment-related items for both groups of teachers and for individual teachers. We also examined longitudinal trends in each group's and each individual teacher's student survey ratings in relation to its data packet score trends over the 3 years.

Analysis of Teacher Surveys

At the end of each year of the study, participants were asked to complete an online survey that addressed their demographic characteristics, education and training, teaching contexts, and participation in professional development opportunities. They were also asked to rate the importance of a list of 24 classroom teaching practices, to report how much emphasis they gave to various learning objectives, and to report how frequently they implemented a variety of teaching practices and assessments with their students. These items corresponded to analogous student survey items and data packet rubric dimensions. Because of the low numbers of teachers completing the online teacher survey, we do not include the results in this article.

Case Analyses of Selected Teachers

Six of the teachers in the study (3 in the National Board group and 3 in the non–National Board group) were selected for in-depth analyses of their teaching practice. All 3 National Board teachers selected showed strong or moderate evidence of growth in the data packet rubric scores from the first to the third year. Two of the non–National Board participants selected also

demonstrated growth in their data packet scores, and 1 did not show much evidence of growth from the first to the third year. The 6 teachers were interviewed using a structured interview protocol that also included an examination of teachers' first data packet from the first year. These interviews and other data collected in the teachers' data packets, teacher surveys, and student surveys from all 3 years were used to write cases of these 6 teachers' formative assessment practices to better illustrate the changes, or lack of change, that occurred.

Findings

The results of the analyses are presented on two main dimensions: differences between the National Board and non–National Board groups and changes over time for the individual participants within each group.

Comparisons of National Board and Non–National Board Groups

We found consistent trends across all of the sources of data—classroom artifacts, student surveys, teacher surveys, and teacher interviews—illustrating a substantial increase in formative assessment practices for National Board candidates as they experienced the certification process, which was maintained in the subsequent year, and significantly greater gains than those experienced by the non–National Board group. Teachers' individual trajectories illustrated how access to professional development opportunities, especially for 2 in the non–National Board group, also made a difference in practice, as did personal situations, such as illnesses, for some teachers. These variations notwithstanding, the overall trends suggested a strong influence of the National Board Certification process and similar professional development activities on teachers' thinking and practice.

Classroom data. Analyses of the data packet scores indicated that teachers in the National Board group began with data packet scores that were lower than the scores of the teachers in the non–National Board group, with overall mean scores of 2.62 and 2.90, respectively. In this first year, none of the teachers in either group had experience with the National Board Certification process. The group of teachers who were eventually identified as National Board candidates in the second year started with mean scores on each of the six dimensions of assessment that were lower than the mean scores of the non–National Board group. However, none of the group means in the first year are significantly different than the total population mean, based on a two-tailed z test.

During the certification year (the second year of the study), the mean data packet scores of the National Board group surpassed those of the non–National Board group. Analysis of score differences between the first and second year indicate that the National Board group had large score gains on all six assessment dimensions, resulting in significantly higher scores than the

Table 3
Mean Data Packet Scores and Mean Score Differences by Group

Dimension	Mean Data Packet Score			Mean Score Differences	
	Year 1	Year 2	Year 3	Year 1 to Year 2	Year 1 to Year 3
View of assessment					
Non–National Board	2.95	2.97	3.27	0.02	0.32
National Board	2.66	3.48	3.68	0.82*	1.02
Range, quality, and coherence of assessment methods					
Non–National Board	3.02	3.01	3.42	–0.01	0.40
National Board	2.78	4.04*	3.77	1.26*	0.99
Clarity and appropriateness of goals and expectations for learning					
Non–National Board	3.13	3.12	3.34	–0.01	0.21
National Board	3.10	3.61	3.60	0.50	0.50
Opportunities for self-assessment					
Non–National Board	2.33	1.79	2.41	–0.54	0.08
National Board	1.84	2.53	2.52	0.69*	0.67
Modifications to teaching based on assessment information					
Non–National Board	3.31	3.17	3.28	–0.14	–0.03
National Board	2.72	3.56	3.40	0.83*	0.68
Quality and appropriateness of feedback to students					
Non–National Board	2.69	3.01	3.01	0.32	0.32
National Board	2.59	3.46	3.30	0.87	0.71

Note. Discrepancies between mean data packet scores and mean score differences are due to rounding.
*Differences between Non–National Board and National Board teachers are significant ($p < .05$).

non–National Board group on four of the formative assessment rubrics (see Table 3). Gains were substantial for 7 of the 9 National Board teachers (see the appendix for individual and group scores reported for each dimension).

In the post-certification year (the third year of the study), the average data packet scores dipped slightly for the National Board group but remained higher than those of the non–National Board group (see Figure 2). Overall, the National Board group increased its average score from the first to the third year by at least 0.5 points on every rubric dimension and by a full point on Views and Uses of Assessment and on Range, Quality, and Coherence of Assessments. The non–National Board group, on average, did not increase any of its rubric scores by as much as 0.5 points over the 3 years. These results show that teachers in the National Board group improved their formative assessment practices while engaging in the certification process and

largely maintained these assessment practices in the following year, suggesting that they had likely incorporated these changes into their repertoires of practice. According to some theories of learning, we would expect a plateau in performance after a period of accretion of new experiences and information (Rumelhart & Norman, 1978); thus, it seems reasonable that we do not see a continued increase in third-year scores for the National Board group.

We further examined individual score trends to understand the third-year decrease for the National Board group and the increase in that year for the non–National Board group. We looked at data packet scores combined with information gained in the final interviews conducted with 15 of the 16 participants. We found that 2 of the National Board candidates who showed strong gains in their data packet scores during the certification year and a decrease in their scores in the third year had serious health problems during that third year. One was out of school for surgery and was restricted to teaching her middle school science classes from a sitting position when she returned. She noted that she had to omit some of the more active assignments from her curriculum due to these physical restrictions. The teacher with the largest decline in third-year scores (SMH 50) missed part of the school year due to surgery and complicating conditions, which most likely affected his relationships with students, his time and energy for participating in the data collection procedures, and his general instructional practices. If we were to assume this teacher's performance was compromised and remove his scores from the National Board group mean calculation, the group mean would show an increase in the third year (from 3.44 in year 2 to 3.56 in year 3) rather than showing a slight decrease (see the appendix).

In examining why the non–National Board comparison group should show an increase in assessment scores in the third year, we again looked at the individual experiences of the participants. The assessment practices of most teachers in the non–National Board group changed little over the course of the 3 years, with the exception of 2 teachers', 1 (MFM 17) who had substantial gains across all of the rubric dimensions, and the other (SFM 56) who had moderate gains. Two-thirds of the first- to third-year gains for the non–National Board group are accounted for by the data packet scores of MFM 17. If this teacher's scores were to be removed from the non–National Board group, the group gains would drop from 0.22 points from year 1 to year 3 to 0.07 points. This non–National Board participant had experienced intensive professional development activities that focused attention on student learning during instruction and addressed teachers' assessment practices. Additionally, the individual score profile of MFM 17 across the 3 years of the study resembles the National Board group score profile when the compromised teacher (SMH 50) scores are removed from the National Board group profile. We discuss these experiences further below.

The other non–National Board participants reported a variety of influences on their teaching practice, but these changes did not show up strongly on our measures of assessment practices. One teacher in the non–National

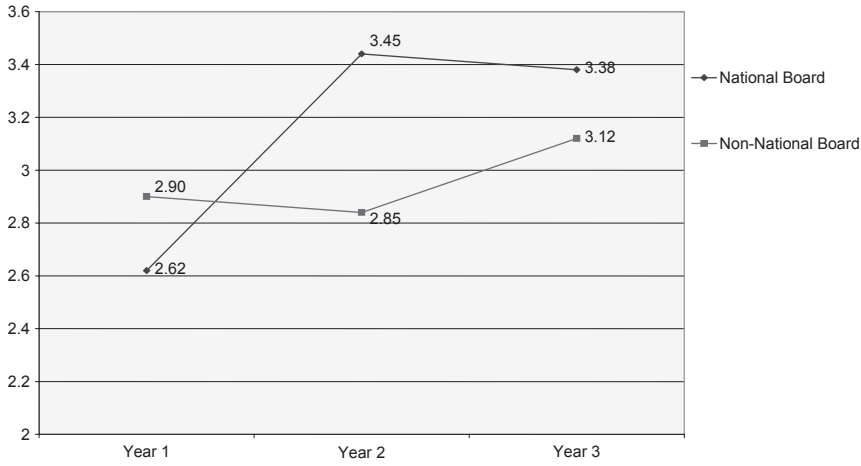

Figure 2. **Data packet scores, years 1 to 3.**

Board group whose scores showed modest gains across the 3 years reported that she had participated in a science inquiry professional development experience at a well-known science center and in district-sponsored gifted and talented training workshops. The other teachers in the non–National Board group with little or no gains reported changes in curriculum and increased use of technology as influencing their classroom practice. One teacher reported participation in a district mentor-training program, but she did not report that this influenced her own practice.

We also found that teachers in both groups felt that their participation in this study had influenced their teaching practices. While the teachers were not aware of the focus on classroom assessment practices in the study, videotaping themselves teaching, recording information such as learning goals for a lesson and their expected outcomes, and feeling as if they were "reporting" to an external agency all gave them incentive to reconsider their teaching practices and ask themselves about the choices they were making.

Student surveys. Analyses of the student survey responses on both closed and free-response questions showed patterns in ratings that were consistent with trends in the data packet scores. Correlation analyses indicate that there were low to moderate, statistically significant relationships between the sum of the ratings in each of the categories of the student survey and corresponding data packet scores. In particular, there were modest relationships between the survey items and data packet scores assessing Modifications to Teaching (r = .55) and between the survey items and data packet scores examining Opportunities for Self-Assessment (r = .47). The survey items in

the Range, Quality, and Coherence of Assessment Methods dimension were also significantly correlated with four of the data packet rubric scores: Views and Uses of Assessment; Range, Quality, and Coherence of Assessment Methods; Clarity and Appropriateness of Goals; and Quality and Appropriateness of Feedback.

We compared year-to-year trends in average student ratings for each group on clusters of survey items using ANOVA. As Table 4 shows, on average, the student ratings of teachers in the National Board group changed significantly on more items from year 1 than the student ratings of teachers in the non–National Board group. Across the 24 survey items, National Board candidates on average were rated significantly higher on nine more items in the second year than in the first year and on six more items in the third. (We reverse-scored the item on frequency of multiple choice tests and quizzes, because this testing format provides little opportunity to gain information about the nature of student thinking.)

National Board candidates as a group were rated more favorably by their students in the second and third years of the study on areas that are more often associated with improved use of assessments for formative purposes, including providing opportunities for self-assessment and for students to reflect on their learning, emphasizing questions that require explanations during class discussion (not only yes or no answers), having students answer questions about what they know before a unit begins (pre-assessment), and engaging in hands-on activities, written reports, and group discussions. Significantly lower ratings on one item indicated a decreased emphasis on giving presentations to the class.

Relative to the first year, non–National Board participants were rated significantly higher on only three items in the second year and lower on one: use of multiple-choice tests and quizzes, which is reverse-scored, increased in this group in the second year of the study. Their students reported increased opportunities to develop data analysis skills and apply science or mathematics to real-world problems and situations. Even stronger gains in the third year, with students noting greater opportunities to do hands-on work, to work on individual or group projects, and to engage in self-assessment, were driven primarily by changes in the student ratings of 1 teacher who had received intensive professional development during the course of the study.

There was a high level of agreement between the first- and third-year changes in the data packet scores and overall changes in the student ratings of teachers. The 4 teachers who had large score gains (average rubric score differences of about 1.0 point or more) in the National Board group and the 1 teacher who had a similar gain in the non–National Board group also had the highest overall number of positive gains in their student survey ratings. In free responses, students of National Board candidates were more likely to report over the 3 years of the study that they knew the learning goals of a unit based on the teacher's instruction or work assigned. In addition, the students of National Board candidates were more likely to report the explicit identification of learning goals through written forms and goal identification

Table 4
Student Surveys 2004–2006, Year-to-Year Mean Ratings on Assessment Items

About How Much Emphasis Is Placed on Each of the Following in This Class? (1 = None, 2 = Little, 3 = Some, 4 = A Lot)	National Board Group			Non–National Board Group		
	Year 1 (n = 366)	Year 2 (n = 360)	Year 3 (n = 371)	Year 1 (n = 290)	Year 2 (n = 249)	Year 3 (n = 280)
Focus on grades	3.77	3.69	3.64	3.63	3.64	3.64
Range, quality, and coherence of assessment methods						
Individual/group project	3.14	3.15	3.23	3.01	3.15	3.18[a]
Homework	3.36	3.42	3.49[a]	3.36	3.39	3.43
Class presentations	2.45	2.31[b]	2.40	2.36	2.37	2.20[b]
Multiple choice test/quiz (reverse-scored item)	3.23	3.05[a]	3.11	2.84	3.02[b]	2.84[b]
Hands-on activity	3.01	3.16[a]	3.11	2.63	2.68	2.80[a]
Portfolio	2.23	2.11	2.18	2.10	2.17	2.19
Notebook/journal	2.85	2.97	2.92	3.29	3.40	3.40
Written report	2.33	2.48[a]	2.48	2.48	2.39	2.35
Clarity and appropriateness of goals and expectations for learning						
Facts/vocabulary	2.91	3.02	3.03	3.09	3.15	3.18
Big ideas/concepts	3.05	3.23[a]	3.20[a]	3.10	3.13	3.08
Skills/procedures to solve problems	2.97	3.11	3.04	2.86	2.89	2.92
Communicate ideas	2.65	2.72	2.73	2.57	2.71	2.60
Logical thinking skills	2.73	2.84	2.81	2.79	2.84	2.86
Data analysis skills	2.81	2.86	2.90	2.61	2.85[a]	2.82[a]
Apply science/math to real world	2.65	2.82	2.78	2.38	2.66[a]	2.58[a]
Self-assessment						
Reflecting on learning	2.75	2.99[a]	2.87	2.71	2.82	2.81
Self-assessment	2.25	2.24	2.44[a]	2.25	2.33	2.40[a]
Peer assessment	2.25	2.29	2.38	2.19	2.56[a]	2.32[b]
Discussing work in groups	2.84	3.03[a]	3.18[a]	2.47	2.61	2.61
Modifications						
Class discussions	2.71	3.04[a]	3.02[a]	3.10	3.02	2.99
Explaining answers	2.87	3.09[a]	3.00[a]	2.90	2.95	2.90
Pre-assessment	2.37	2.54[a]	2.37[a]	2.37	2.24	2.31
Feedback						
One-on-one conference	1.79	1.76	1.74	1.84	1.93	1.94
Items with a significantly higher rating than in year 1 (2004).		9	6		3	5
Items with a significantly lower rating than in year 1 (2004).		1			1	1

[a]Items with a significantly higher rating than in year 1 (2004).
[b]Items with a significantly lower rating than in year 1 (2004).

in the second year of the study, when their teachers experienced the National Board Certification process.

Although few of the open-ended responses specifically addressed teacher feedback, the students of Sam, a National Board candidate highlighted in the case reported below, frequently reported that he provided feedback aimed at improvement of work. This teacher also posted one of the highest score changes on the Quality and Appropriateness of Feedback rubric in his data packets. The students' comments from this class were very articulate about the regular opportunity that they had to revise their work, as these typical responses indicate, "My teacher and I talk about revisions I need to do for my work," and "If there is something wrong with our work, we will revise it until it is right." In contrast, students of Helen, a non–National Board participant, reported that individual feedback on notebook checks consisted primarily of a notation of points earned and a comment about needing to improve the neatness of their work.

Cases of Teacher Change

We developed case studies of 6 individual teachers in order to further understand the variety of ways that assessment practices were enacted and changed over time as well as the complex factors that influenced the teachers' development. We selected the case teachers to represent the range of teaching contexts and changes in practices within each group. From the National Board group, we selected 2 teachers who had substantial gains in their scores over the 3 years—Sam and Marie, who taught science in a low-performing high school and high-performing middle school, respectively—and 1 high school mathematics teacher, Isaac, who showed moderate gains in his data packet scores. From the non–National Board group, we selected Iris, the high school mathematics teacher in a high-performing school who showed the strongest gains in this group; Helen, a middle school science teacher in a mid-performing school who showed small gains; and Louis, a science teacher in a private high school who showed little or no gain across most dimensions of the assessment rubrics.

To construct each teacher's story of change, data from the data packets submitted by the teachers across the 3 years were examined, including videotapes, lesson plans, and student work. We also conducted final interviews that probed specific questions related to assessment practices, the teachers' perceived views of change over time, and influences on teaching practices. Each case was written independently by a member of the research team and reviewed by the entire team in light of the data before being finalized.

We briefly present the case of Sam to provide a holistic view of how assessment practices changed for a teacher in a low-performing school while he was engaged in the National Board process. We then provide an analysis across the six cases to illustrate more broadly the ways teachers viewed assessment and the variety of factors that influenced their practice during this study.

Sam's changing classroom assessment practices. Sam, who had been teaching for 5 years at the beginning of the study, was a 33-year-old, White, male teacher. He was credentialed to teach secondary science and held a master's degree in science education. The school where he taught had a long history of challenges and poor performance on standardized measures of achievement. During this study, the school's academic performance, as measured by state-administered tests, ranked in the bottom decile statewide. Sam's rubric scores increased each year of the study, starting with scores at the 2 and 3 rubric levels and increasing each year to the 4 and 5 rubric levels in the third year.

From the beginning of this study, Sam was strongly aware of the state standards for science, and he deliberately referenced the standards in his teaching orally and through posters in his classroom. We viewed this practice as a means by which Sam communicated the goals for learning to his students on a regular basis. Over the 3 years of this study, Sam developed ways to reference the state science standards in his teaching, and he began to use these goals to guide his instructional interactions with students.

In the second year, while he was undertaking National Board Certification, we saw that after introducing a new concept, Sam intentionally used open-ended discussion questions to ask students what they understood and what they felt they were still confused about. He embedded "minute papers" in lessons, asking students to write briefly about the ideas from the lesson in their own words as a way for them to self-assess their developing understanding of the key ideas. He also used "take-home quizzes" in which students were asked a series of questions drawn from the day's lesson as a means for the students to gauge their understanding of that lesson. At the end of his second year, Sam indicated in an interview that he had begun to think about how he needed to focus on the "big idea" in his instruction, not just as the stated goal of the lesson. At the time of the final interview, Sam described how he had eliminated many curricular activities and assessments "that don't allow students to show the most important understanding."

Sam's students also seemed to be increasingly aware of how the learning goals shaped Sam's instructional interactions. In the student responses to the open-ended survey question about how they know what the teacher wants them to learn, 11% of the student comments in the first year indicated that the "aims and purposes" of the lessons were made explicit. In the second year, the frequency increased to 60%; it remained at 50% in the third year.

We also saw changes in the way Sam used the assessment information he regularly gathered in his teaching. In the first year of the study, Sam consistently began lessons with activities designed to help him understand what his students already knew. However, it was not evident how he then used that information to choose appropriate activities for his students or to make decisions about how to proceed with instruction. Sam went through the motions of pre-assessing his students' knowledge and understanding, but he did not use this information formatively in his instructional planning. In the third-year teaching videos, we observed Sam in conversations with his students, pushing them to elaborate on

their thinking, to connect it to other concepts, and to apply their understanding in concrete ways. Sam's instructional decisions (what question to ask next, who to call on, and what examples he chose to provide) built on what the students were saying and demonstrating during the instructional interactions. In one of the third-year interviews, Sam discussed how he began the lesson sequence with activities that were designed to uncover what students had mastered and retained from middle school so he could gauge what they already knew, and only then did he begin to set goals for the class.

These changes in Sam's assessment practices can be seen by comparing how he presented the same lesson sequence 2 years apart. In the first year, his assessment plan for the unit on phase changes included only one planned assessment, a quiz, which was the basis of students' grades. Although it was presented as assessing students' beginning knowledge, there was no evidence in his plans of another assessment to assess their later knowledge. In the third year of the study when Sam submitted the same lesson sequence, the lesson plan overview identified five specific assessments related to the learning goal, including a brainstorming activity, a lab report, classroom discussions, and a quiz. In the video from the third year, Sam listened carefully for understanding and for misconceptions when his students defended their answers from a lab exercise and a quiz during a class discussion. He then addressed the students' ideas through questioning and discussion, repeatedly coming back to the students' comments and ideas. Finally, he had students revise their work. In his final interview, Sam acknowledged that during the first year he thought about assessment as being mostly about grades. He said that he now believed that his focus was on learning what students know and do not know and selecting strategies to help students build on their knowledge and explore topics more deeply.

Across the 3 years of the study, Sam also changed in the way he worked with students to revise their work. In the first year, Sam's students were given limited opportunities to revise work to meet learning goals. In a year 1 interview Sam discussed how he used a rubric to help students understand their performance but felt this strategy was not successful. In a year 2 interview he described how he encouraged students to revise and helped them to understand the goal by showing them examples of mastery from former students. In the year 3 teaching videos, Sam was observed engaging in classroom conversations with his students about their revisions, pushing them to elaborate on their thinking, to connect it to other concepts, and to apply their understanding in concrete ways.

By the third year, Sam's feedback on student work was less focused on telling the students what was wrong by making his own corrections on their papers and more focused on clear directions to them about what needed to be done to make the work better. In the final interview, Sam described a more elaborated set of revision guidelines that asked students to identify "what they have done, what they think they need to do, and how they will accomplish those things." He built in opportunities for the students to understand and act on the feedback as part of the assignment. In the open-ended

student survey responses in the third year, students consistently cited the comments and the teachers' encouragement "to revise so they could understand" as the kind of feedback they received from their teacher.

Changes observed across the cases. The classroom assessment practices of the teachers who demonstrated high scores and strong gains on the assessment rubrics for both the National Board candidates and the non–National Board participants can be characterized as becoming better integrated into teachers' ongoing instruction. The teachers began to see assessment opportunities in daily interactions with students in addition to the periodic tests or quizzes that were part of their existing practice. As the teachers' assessment repertoires expanded, their relationships to their students seemed to change, with the teachers attending to their students as learners and developing a stronger appreciation of their role in supporting students' learning, not solely monitoring their grades.

For example, Sam's attention to revision of student work and his increasing attention to handing over responsibility for that revision to the students indicated his willingness to use assessment as an opportunity for students to develop their ideas and skills until they reached a level of understanding that the teacher *and the students* felt was satisfactory. Sam also developed a greater awareness of his learning goals and the big ideas that he wanted his students to learn. While he had already used the state academic standards as a guide for his instruction, we noted a stronger alignment of assessment with instruction, especially in the feedback Sam gave to his students, as he became more explicit in his expectations for student understanding. Sam became more deliberate in planning informal assessment opportunities. This was less a matter of launching new techniques than of becoming more attentive to students during class discussions, focusing his comments to them toward specific learning goals and structuring self-assessment opportunities that allowed the students more agency in their own learning.

Similarly, Marie, another National Board candidate, shifted her classroom practices to integrate assessment in the ongoing instruction in her classroom. This was demonstrated most dramatically in her questioning strategies in classroom discussions. She characterized her prior practice as the "illusion that everybody is just floatin' right along and they're getting it" during class discussion. We observed the evolution toward her current practices in which she and the class "hash it out together" to get at *how* the students understand an idea. Additionally, Marie changed the way she used the class notebooks she expected the students to keep. Where they had once been a chore to grade and received cursory attention, she began to see them as a source of information that she could draw upon. The notebooks became windows into student thinking that provided opportunities for immediate feedback and revision, thus supporting the developing ideas of the students.

The third National Board candidate, Isaac, began his participation in the study with low rubric scores and made modest gains during the three years. His initial practice involved strong adherence to the textbook curriculum,

which provided little opportunity for modification to practices based on students' learning needs. Isaac chose to use one-on-one work with students when they requested it to help them with their individual needs. The feedback that he provided consisted of additional explanations for working the mathematics problems during independent work time and was focused on leading students to the correct answer. Isaac's new self-assessment strategies, such as practice tests, pre-tests, and students' monitoring of their own grade over time, began to involve students in the assessment processes of the classroom.

Iris, the non–National Board participant who demonstrated strong gains in her assessment rubric scores, also demonstrated an increasing awareness of the opportunities she had to gauge student understanding through everyday interactions. She participated in intensive professional development, sponsored by the Noyce Foundation, anchored around authentic assessments in mathematics, which drew her to deeper understanding of student learning. As Iris moved away from large-scale final assessment projects, she moved toward spending time on instructional activities that required her students to work through problems and come up with solutions. She provided more opportunities for her students to have conversations in class through "Think, pair, share" activities, and she orchestrated instruction in a class discussion format more often. This allowed her to pause and ask students if they understood and to more fluidly redirect the lesson to issues that seemed to be problematic for the students.

Louis, a non–National Board participant, began and ended his participation in this study with assessment practices that scored at the 3 and 4 level on the assessment rubrics. Even though he did not register much gain in scores, and actually had a drop in his scores in the second year while he took on administrative responsibilities for the school, we saw attention to student learning in his assessment practices as well as an integration of these practices in his daily instruction. Louis wove the learning goals into his instruction in multiple ways—returning to them in class discussion, in regular self-assessment opportunities, and in a yearlong portfolio record of progress toward the school and course competencies. We saw Louis's students actively questioning their own understandings in class discussions, and we saw Louis using a wide array of assessment strategies that provided both him and his students with opportunities to wrestle with and demonstrate their understanding.

Helen, a non–National Board participant, showed small gains in rubric scores but remained on the low end of the rubric scale. Her assessment practices focused on student work as a record of production for grading and reporting purposes. Even though she struggled outwardly with important questions of how to help students better understand the substance of her course and how to monitor their learning better, she had nothing in her repertoire to help her make the improvements she desired. She made some gains in modifying her instruction based on what she was observing in her students' learning, possibly as a result of her participation in a professional development opportunity that focused on science as inquiry.

In summary, the National Board candidates in these cases identified changes in their conceptions of assessment as shifting from a focus on grading to the use of assessment for formative purposes. Hand in hand with that shift was a movement away from teaching for discrete facts to teaching for conceptual understanding and aligning those assessments better with learning goals. Marie reported

> a shift from my need as a teacher to assess, meaning I need to have a grade in my grade book so I can communicate a grade on the report card to assessing so that the students can learn something better.

Isaac reported that he is now less interested in using assessments to assign grades, that he tries to make his exams more of a learning activity, and that assessment is a feedback process for him: "If the assessment results are crummy, there is still something I should do to improve in my teaching."

The non–National Board participant who demonstrated scores at the 4 and 5 level of our measures of formative assessment similarly indicated an understanding of classroom assessment as residing in the interactions with the students through classroom discussions, feedback loops, and students' engagement in talking and thinking about their own learning. In the non–National Board case of low performance on our measures of formative assessment, the classroom assessment practices were primarily focused on keeping records of student work production. Even though the teacher knew that something was absent from her instructional interactions with the students, she did not demonstrate to us an understanding of the role that assessment might play in supporting the development of her students' understanding. And while she seemed to desire improvement in her teaching, she did not have professional resources or processes for introducing changes into her practice.

Accounting for Changing Classroom Assessment Practices

Across the final interviews with study participants, we identified four main factors that they reported as influencing change in their classroom practices: the National Board Certification process, other formal professional development opportunities, collegial interactions among teachers, and participation in the research study itself.

National Board Certification. When asked in an open-ended question to describe what prompted changes in their teaching practices, 6 of the 8 National Board candidates who participated in the final interviews[1] cited the National Board Certification process as a key factor that led to changes in their teaching practices. National Board candidates indicated a variety of aspects of the National Board Certification process that catalyzed change in their classroom assessment practices. Teachers reported that the teaching standards of the National Board provided a set of clear goals for practice and a practical sense of what constituted those goals. The standard for "assessment"

defines assessment practice as the collection of information from a variety of sources and the use of that information to inform teachers' instructional decisions. Introducing this definition of assessment to National Board candidates who do not already hold that view often changed teachers' operating definitions of assessment and broadened their practices. One teacher voiced an experience common among the National Board candidates, describing how his practice had changed from a focus on tests to a more comprehensive portfolio including tests, projects, other work samples, and discussions. He reported that this variety of assessment methods allowed him to provide more immediate and comprehensive feedback to students.

Teachers also reported that the portfolio entries required them to engage in practices that brought assessment practices into sharper focus, helping them see how assessment operated in their classrooms. One teacher noted that the portfolio prompts consistently asked for evidence of learning and promoted specific, systematic attention to student learning that helped her develop tools to evaluate her teaching practices and students' learning. This emphasis on evidence was critical for her, with every prompt in the portfolio directions coming back to student learning. She reported that this emphasis helped her to attend to "hard evidence" versus anecdotal evidence of student learning. Another commented that the process "definitely forced me to slow down and to look closely at whom I was teaching," pointing specifically to the portfolio entry requiring in-depth examination of two students and their work and progress over time. "I continue to do this now," he stated.

Other impacts of the National Board Certification process on teaching and professional life noted by candidates were a greater emphasis on doing meaningful hands-on activities; greater coherence in lesson units; greater efforts to encourage parental participation; an improved relationship with students and parents related to a less adversarial grading system; promotion of the sharing of professional learning with the larger teacher community; a stronger stance toward reflecting on teaching; a better grasp of content knowledge prompted by the subject matter assessment center exam; and encouragement to continue what they had been doing by feeling valued as professionals.

In addition, in citing what was instrumental about the certification process, all 6 of these teachers cited the value of the collegial interactions that they had with other teachers in their National Board candidate support groups. These teachers appreciated the opportunities to analyze the National Board standards as a group. The teachers in the support group participated in standards-based workshop sessions in which teachers dissected the standards, brainstormed classroom practices that aligned with the standards, and used classroom artifacts to engage in standards-based analyses of teaching. These activities were conducted in both heterogeneous subject-matter and grade-level groups and homogeneous subject-matter groups (e.g., all science teachers together). This provided the candidates both broad-based perspectives on teaching and assessment and subject-matter-specific practices.

These candidates also reported that the collegial critique of their videotaped lessons allowed them to hear other teachers' perspectives on their

practice, with which they were sometimes too intimate to analyze. The collegial video analysis also gave them windows into other teachers' classrooms and thinking. Seeing other possibilities of practice caused some teachers to both adopt new practices and cease some old ones. As one noted of his work with these colleagues, "I saw things in their practice that I wanted in my own, and I saw things in my practice that I didn't want there anymore."

The certification process in conjunction with the support group environment created opportunities for collegial analysis, reflection, and constructive critique of videotaped lessons; sharing of teaching strategies and ideas with teachers from across a variety of schools and districts; and camaraderie with like-minded teachers with similar goals of "trying to improve the profession and our teaching."

Structured professional development. Catalysts for change in classroom assessment practices included not only the National Board Certification process but other opportunities for professional development and collaborations with colleagues as well. Sam identified the Japanese Lesson Study model, Iris reported participating in Noyce Foundation workshops that have a strong focus on student work analysis, and Helen participated in a science inquiry professional development program. These programs share some common features with the National Board Certification process in that they focus on actual classroom practice—the minute-by-minute and the day-to-day actions in the classroom as well as the artifacts that result from classroom practice. These professional development opportunities also share a focus on evidence of student learning embedded in the artifacts and interactions of the classrooms. Finally, these programs share an emphasis on structured reflective opportunities for teachers while re-focusing them on learning goals.

Collegial interactions. In addition to the support group for National Board Certification, teachers identified on-site collegial interactions that influenced their practice. For some teachers, these were formal mentoring programs in which they were observed by colleagues on a regular basis, and for others the opportunities were informal. The teachers reported that these opportunities to work with colleagues allowed them to share ideas and practices, make everyday curriculum decisions, and participate productively in school reform initiatives.

Participation in the study. Finally, teachers' participation in the research study itself likely caused some changes in practices. While the teachers were not aware of the focus on classroom assessment practices in the study, videotaping themselves teaching, recording information such as learning goals for a lesson and their expected outcomes, and feeling as if they were "reporting" to an external agency all gave them incentive to reconsider their teaching practices and ask themselves about the choices they were making. For example, during the final interview, in response to a question as to what had changed in his classroom teaching practice and what prompted those changes, Louis,

a non–National Board participant, immediately began to discuss the impact participation in the study had on him:

> [I became] more conscious of what I was doing because I was going to report it. Just being more conscious is what prompted the changes. I've done more videotaping that hasn't been required for us just because it gives me some good feedback, and also to show students and have them evaluate their own behaviors from the videos. [I've been] more aware of what I'm doing, because the whole accountability issue is important: "What am I doing? Why am I doing it?" . . . and reporting to someone. . . . In other words, "Tell us what you're doing that's worthwhile."

To the extent that participation in the study improved the teaching and assessment practices of non–National Board teachers, as some reported it did, it provided a partial "treatment" that likely reduced the measured differences between National Board and non–National Board teachers' changes in practice.

An Integrated Look at Influences on Practice

Taken together, it appears that teachers' classroom teaching practices can be influenced by professional activities that allow them the opportunity to closely examine their own practice. Whether through the National Board Certification process (the critical case under exploration in this study), the professional development provided by other organizations with similar characteristics, or the intervening data collection procedures of this study, the teachers were afforded the opportunity to develop their teaching repertoires by starting with an analysis of their own practice.

The teachers who experienced the National Board certification process reported that the requirements of analyzing their classroom practice with a focus on assessment as defined by the National Board teaching standards introduced them to new ways of viewing the role that assessment plays in their everyday instructional interactions. The process of videotaping their teaching and analyzing it also brought elements of their practice into sharper focus. Similarly, the teachers who experienced professional development opportunities sponsored by other organizations that used processes of reflecting on artifacts from classroom teaching reported that they used assessment information to understand student learning and to gauge their instructional decisions. And while we attempted to design a set of data collection procedures that would not imitate the analysis and reflection procedures of the National Board Certification process, teachers reported that our procedures of videotaping their practice and reporting on a few open-ended questions was enough intervention to prompt them toward reflecting on the value and purpose of their instructional strategies.

While analysis of one's own teaching practices served as a starting point, the direction for improvement seems to have been guided through collegial

interactions and by the goals of the programs in which the teachers participated. From the interactions with others, the teachers in the National Board group gained a more nuanced understanding of the National Board assessment standard in practice, saw a variety of other practices, and engaged in critique of their and other teachers' practice. The collegial opportunities for teachers in the non–National Board group also seemed to support the development of practice through school-based reform efforts based on a vision of instruction that supports student learning.

This study was particularly interested in effects on classroom assessment practices, given the strong correlations that have been demonstrated between assessment that supports learning and learning outcomes. It seems that the professional development activities that had an explicit focus on assessment (the National Board teaching standards for assessment and the Noyce Foundation's program focus on assessing student work) were linked with improvement on our measures of classroom assessment practices. While other influences such as curricular changes, technology, and mentor training were identified by the teachers who did not show gains on our measures, these interventions may have had impacts on dimensions of practice that were not the focus of this study.

Conclusions and Implications

Given our small sample size and the non-random assignment to candidate to comparison groups, the findings of this exploratory study are a step toward greater understanding of what influences change in teachers' classroom practice. We provide a starting point for future research by examining the National Board Certification process as a critical case of professional development.

This study suggests that the certification process, and similar professional development opportunities, may provide a fruitful avenue for further research. Although as a group, the National Board candidates taught in schools with more disadvantaged student populations and fewer instructional resources, they appeared to experience greater advances in their teaching practices associated with formative assessment and scored higher than a comparison group by the third year of the study. The candidates attributed most of the changes in their practice to the National Board Certification process. The only teacher from the non–National Board group to experience comparable gains experienced an intensive professional development experience very similar in many ways to the National Board Certification process. The findings suggest, then, that professional development strategies like those provided by National Board Certification may help to change teachers' formative assessment practices and, as we saw, their instruction more generally.

In considering directions for further research, it is useful to examine some of the research and theory that illuminate how these influences may operate, so that more formal tests of these ideas may be developed. Research on formative assessment has gained much momentum in recent years as

more data indicate that classroom assessment practices can lead to improved student achievement and can reduce the achievement gap (Stiggins & Chappuis, 2005). Cognitive research also points to the importance of instruction that takes into account students' current understanding and provides ongoing support of student learning through feedback that guides revision (Gardner, 2006). Thus, work that guides teachers to develop such practices may be particularly generative of changes in practice because teachers may see improvements in student learning that encourage them to continue to develop their skills in this area.

Ingvarson (1998) suggests that a system of professional development grounded in a set of professionally defined standards is potentially effective at changing practice because it offers a clear vision of what teachers should be getting better at doing in the classroom. Professional standards "provide goals for professional development that constitute a stable, challenging, and long-term agenda for professional development" (p. 130). The National Board Certification process offers teachers an opportunity to engage in reflection and analysis of their teaching practices using rigorous standards as "tools for critique" (p. 137), as well as valid performance assessments that can guide teachers as they seek to enact the standards in specific classroom practices and give them feedback about what they are doing and how well.

Other researchers have also identified features of effective professional development models that align with features of the National Board Certification process (Hawley & Valli, 1999; National Staff Development Council, 2001). Chittenden and Jones (1997) identify five components of the Board Certification experience that reflect these features and seem critical in their research to enhancing teachers' professional knowledge and skills:

- a framework that provides a vision of good teaching and that has heuristic value for critical analysis,
- a process for grounding the abstractions from the framework in the realities and evidence from daily classroom life through documentation,
- systematic work with colleagues through collaboration,
- allocating specific time for regular meetings with colleagues to keep the demanding process on the teachers' agenda despite the ongoing pressures and demands of teaching, and
- both formative and summative evaluation with commonly understood criteria for evaluation that introduce a degree of accountability (pp. 16–17).

Finally, the National Board Certification process allows professional development to be a personal process for the teacher (Sato, 2003), capitalizing on reflective opportunities for teachers to hypothesize about what is working well and what needs to be improved in their own practice rather than being asked to adopt wholesale new models of teaching that are disconnected from their daily contexts.

Teachers in our study cited these features as influences on their teaching practices. The teaching standards were an important driver of the change,

and the standard for assessment provided a vision of practice that most of the teachers did not have prior to their engagement with the National Board Certification process. Systematic analysis and reflection on their classroom work was prompted by questions that focused on evidence of student learning and alignment of practice with the teaching standards. The teachers in this study also reported that their participation in a collegial support environment in which teachers shared ideas, examined each other's teaching, and jointly critiqued each other's work played an important part in their understanding of the standards enacted in practice.

For this study, we constructed rubrics for six dimensions of formative assessment as a means of identifying, describing, and measuring classroom practices. Using these measures of formative assessment, we were able to note specific elements of practice that aligned with the literature on formative assessment and identify changes over time in teachers' practice based on videotapes, student work samples, and interviews with the teachers about their instructional intentions. Within our research team we achieved strong inter-rater reliability in using these rubrics to measure teachers' classroom practice. We have already seen the value of articulating these dimensions of formative assessment as study groups of pre-service and practicing teachers have begun to use these rubrics as self-assessment tools. Some professional development programs are also beginning to use these rubrics to help teachers better understand the purpose and intent of formative assessment while providing a vision of what strategies might look like. Further research with larger numbers of teachers could help evaluate and strengthen the validity of these measures and test the generalizability of our findings to broader samples.

In addition to larger and better controlled studies, further research might seek to tease apart some of these factors as more or less influential on practice and explore the conditions under which they may be influential, including the kind of supports and expectations that need to be present to gain value from the use of reflective, collegial work around standards. Other enduring issues in studying the effects of professional development opportunities for teachers are the degree of the impact on practice and the duration of changes. This study attempted to explore how changes in practice are sustained after the intervention by examining practice in the post-certification year. Extending the study over 3 years gave us a longer view of change and a sense of the trajectories of teacher practice; however, it simultaneously made teacher retention in the study more difficult. These are issues that need to be addressed in future longitudinal designs, perhaps with greater incentives for teachers and less burden on them for the assembly of classroom data to be analyzed. Finally, future research should examine what kind of teaching conditions and supports are needed for teachers to sustain practices that are learned in intense professional development experiences.

APPENDIX
DATA PACKET SCORES FOR NATIONAL BOARD AND NON–NATIONAL BOARD GROUPS, YEAR 1, YEAR 2, AND YEAR 3

	Views and Uses			Range, Quality, Coherence			Clarity and Appropriateness of Goals			Opportunities for Self-Assessment			Modifications to Teaching			Quality and Appropriateness of Feedback			Mean Rubric Score		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
MMH02	1.44	2.25	2.75	2.00	2.63	3.00	2.13	2.13	3.00	1.06	1.44	1.88	1.31	1.88	2.75	1.19	2.00	2.50	1.52	2.05	2.65
SFM18[a]	2.38	3.94	3.38	2.63	4.00	3.31	2.69	3.94	3.25	1.44	2.00	2.75	2.63	3.31	2.88	1.75	3.00	2.56	2.25	3.36	3.02
SFH20	3.69	3.44	3.50	3.50	3.69	3.50	4.06	3.31	3.88	2.44	1.88	1.88	3.63	3.50	3.75	3.44	2.88	3.50	3.46	3.11	3.33
CFM26	2.13	4.00	4.63	2.06	4.25	4.50	2.82	3.75	4.13	1.25	2.50	3.25	2.63	4.50	4.25	2.63	4.25	3.31	2.25	3.88	4.01
SMH41[b]	3.88	4.06	3.88	4.13	4.63	4.38	3.88	3.88	3.63	3.38	3.06	3.25	3.44	3.31	3.25	3.88	4.38	4.31	3.76	3.89	3.78
SMH48	3.13	3.75	4.88	3.00	4.50	4.50	3.38	4.13	4.00	2.00	4.00	3.63	3.44	4.25	5.00	3.13	4.25	4.88	3.01	4.15	4.48
SFH 49	2.75	3.00	3.88	2.92	3.94	3.88	3.29	3.31	3.63	1.00	2.25	1.00	2.67	3.38	3.25	3.00	3.13	3.63	2.60	3.17	3.21
SMH50[c]	1.56	3.13	2.00	1.81	4.00	2.88	2.38	3.88	2.88	1.00	1.63	1.00	1.69	3.25	1.75	1.38	3.13	1.25	1.64	3.17	1.96
SFM63	3.00	3.75	4.25	3.00	4.75	4.00	3.31	4.13	4.00	3.00	4.00	4.00	3.06	4.63	3.75	2.94	4.13	3.75	3.05	4.23	3.96
National Board group mean	2.66	3.48	3.68	2.78	4.04	3.77	3.10	3.61	3.60	1.84	2.53	2.52	2.72	3.56	3.40	2.59	3.46	3.30	2.62	3.45	3.38
MFH15[d]	2.83	—	3.00	2.50	—	2.88	2.58	—	2.50	2.54	—	2.75	3.08	—	3.50	1.88	—	2.31	2.57	—	2.82
MFM17[e]	2.31	3.38	3.50	2.56	3.25	3.63	2.81	3.38	3.69	1.19	2.25	2.56	3.06	3.00	3.81	1.94	3.63	3.38	2.31	3.15	3.43
SFM37[f]	3.25	3.88	3.94	3.50	4.06	4.06	3.31	3.94	3.88	3.31	2.44	2.00	3.44	3.94	3.63	3.38	4.00	4.00	3.36	3.71	3.58
SMH42	3.69	3.00	3.63	4.04	3.00	4.13	3.75	3.00	3.81	3.19	1.75	2.56	3.94	3.75	3.19	3.50	2.25	3.56	3.68	2.79	3.48
SMH47	3.00	2.31	3.00	3.00	2.56	3.00	4.00	2.75	3.00	1.00	1.06	2.00	4.00	2.88	2.25	3.00	2.63	2.25	3.00	2.36	2.58
SFM54	2.06	2.13	2.63	2.19	2.06	2.69	2.06	2.38	3.00	2.00	1.56	2.63	2.25	1.69	2.94	2.06	1.75	2.63	2.10	1.93	2.75
SFM56[g]	3.50	3.13	3.19	3.38	3.13	3.56	3.38	3.25	3.50	3.06	1.69	2.38	3.38	3.75	3.63	3.06	3.81	2.94	3.29	3.13	3.20
Non–National Board group mean	2.95	2.97	3.27	3.02	3.01	3.42	3.13	3.12	3.34	2.33	1.79	2.41	3.31	3.17	3.28	2.69	3.01	3.01	2.90	2.85	3.12

Note. Average scores represent the average of two data packet scores for each year. Scores within years were relatively stable. Dashes indicate that no data were available.

[a]Surgery in year 3 limited the teacher's physical mobility in the classroom.

[b]The teacher began the National Board Certification process in year 2 and was fully engaged before deciding to withdraw his candidacy. He remained part of the National Board group due to his substantive involvement in the support group and full exposure to the certification process.

[c]Surgery in year 3 resulted in the teacher's missing part of the school year.

[d]The teacher did not submit data packets in year 2 but was kept in the study for year 1 to year 3 comparisons.

[e]In year 3, the teacher participated in professional development focused on modifying assessment practices to accommodate students' learning needs.

[f]The teacher participated in the study without exposure to National Board Certification in year 1 and year 2. The teacher began National Board candidacy in year 3 and did not begin working on National Board portfolio until after submitting year 3 data packets for the study in January 2006.

[g]In the summer before year 3, the teacher participated in professional development focused on analyzing student learning based on analysis of student work.

Note

This project would not have been possible without the talent, hard work, and long hours of a group of researchers from the Stanford University School of Education who participated in the study design, data collection, and data analysis. We extend much gratitude to J. Myron Atkin, Vicki Baker, Sandra Dean, Eric Greenwald, Maria E. Hyler, Gloria I. Miller, Ixchel Samson-Adamek, and Tseh-sien Kelly Vaughn. We are also indebted to the teachers who participated in the study for their time and attention to our data collection protocols. We also thank the anonymous reviewers and the editors who carefully read and responded to our work. This study was supported by a research grant from the National Board for Professional Teaching Standards. Opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the National Board for Professional Teaching Standards or the authors' universities.

[1]Of the 9 National Board candidates, 1 was not available to participate in the final interviews.

References

Athanases, S. Z. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *Elementary School Journal, 94*(4), 421–439.

Baratz-Snowden, J. (1990). The NBPTS begins its research and development program. *Educational Researcher, 19*(6), 19–24.

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

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation at the University of North Carolina at Greensboro.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474–482.

Cavaluzzo, L. C. (2004). *Is National Board Certification an effective signal of teacher quality?* (National Science Foundation No. REC-0107014). Alexandria, VA: CNA Corporation.

Chittenden, E., & Jones, J. (1997, April). *An observational study of National Board candidates as they progress through the certification process*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4*(1), 37–73.

Gardner, J. (Ed.). (2006). *Assessment and learning*. London: Sage.

Goldhaber, D., & Anthony, E. (2005). *Can teacher quality be effectively assessed?* Seattle: University of Washington and the Urban Institute.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook for policy and practice* (pp. 127–150). San Francisco: Jossey-Bass.

Ingvarson, L. (1998). Professional development as the pursuit of professional standards: The standards-based professional development system. *Teaching and Teacher Education, 14*(1), 127–140.

Lustick, D., & Sykes, G. (2006). National Board Certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5). Retrieved March 1, 2006, from http://epaa.asu.edu/epaa/v14n5/

National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded version; J. D. Bransford, A. L. Brown, & R. R. Cocking, Eds.). Washington, DC: National Academy Press.

National Staff Development Council. (2001). *Standards for staff development* (Rev. ed.). Oxford, OH: Author.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension monitoring activities. *Cognition and Instruction, 1*(2), 117–175.

Rumelhart, D., & Norman, D. (1978). Accretion, tuning and restructuring: Three modes of learning. In J. W. Cotton & R. Klatzky (Eds.), *Semantic factors in cognition* (pp. 37–53). Hillsdale, NJ: Lawrence Erlbaum.

Sato, M. (2000, April). *The National Board for Professional Teaching Standards: Teacher learning through the assessment process*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Sato, M. (2003). Working with teachers in assessment-related professional development. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 109–120). Arlington, VA: National Science Teachers' Association Press.

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science, 8,* 173–190.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). *An examination of the relationship of the depth of student learning and National Board certification status*. Boone, NC: Office for Research on Teaching, Appalachian State University. Retrieved June 1, 2006, from http://www.nbpts.org/UserFiles/File/Applachian_State_Study_Smith.pdf

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Merrill.

Stiggins, R. J., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice, 44*(1), 11–18.

Tracz, S. M., Sienty, S., & Mata, S. (1994, February). *The self-reflection of teachers compiling portfolios for national certification: Work in progress*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Chicago.

Tracz, S. M., Sienty, S., Todorov, K., Snyder, J., Takashima, B., Pensabene, R., Olsen, B., Pauls, L., & Sork, J. (1995, April). *Improvement in teaching skills: Perspectives from National Board for Professional Teaching Standards field test network candidates*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 117.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: American Association for Curriculum Development.

Manuscript received May 21, 2007
Revision received December 25, 2007
Accepted February 3, 2008