Early Adolescence/English Language Arts
Classroom Observation Study

CSE Technical Report 433

Robert Land
Center for the Study of Evaluation
UCLA

July 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-6511
(310) 206-1532

The material contained herein is based upon work supported by the National Board of Professional Teaching Standards through a subcontract with the University of North Carolina at Greensboro. Any opinions, findings, conclusions, and recommendations expressed herein are those of the author(s) and do not necessarily reflect the views of the National Board for Professional Teaching Standards or the University of North Carolina at Greensboro.

EXECUTIVE SUMMARY

The Technical Analysis Group of the National Board for Professional Teaching Standards (NBPTS) contracted with the Center for the Study of Evaluation (CSE) to develop and validate a classroom observation system for possible use in the certification process for accomplished teachers. The present report reflects results of a study dramatically down-scaled from that originally proposed. The major research questions addressed in this study involve the reliability and cost of classroom observation and the extent to which observation scores are consistent with candidate certification decisions.

CSE developed and piloted an observation system that is based in the Early Adolescence/English Language Arts (EA/ELA) Standards for National Board Certification. We trained observers in the use of the system, and, in the fall of 1994, we observed and scored the teaching of 15 candidates for EA/ELA certification in 1993-94. We analyzed scores for reliability and for the degree to which they were grounded in the EA/ELA Standards. We compared candidates' scores on observation with their scores on the EA/ELA exercises. We also interviewed the teachers, asking them for their perception of the validity of classroom observation.

Following are the major findings of this study:

- The Classroom Observation System as designed and implemented provides reasonably reliable scores. Pairs of observers independently scored 90 separate classes, six for each of the 15 teachers. Observers used a 4-point scale and were permitted to modify ordinal (1, 2, 3, 4) scores with pluses or minuses. Observers gave the same ordinal score 70% of the time. Generalizability coefficients for individual days of observation with two raters ranged from 0.79 to 0.94. Generalizability across the three days was lower, 0.57, because of day-to-day variations in teachers' scores.

- The Classroom Observation System is adequately grounded in the Standards. Using lists of indicators of accomplished teaching directly derived from the Standards, observers were asked to list their reasons for assigning a particular score. On average, observers cited just under six reasons (out of 20 possible), and about half of the reasons cited by independent observers of the same class matched.

- If observation were the only measure of accomplished teaching, achieving high levels of generalizability would be prohibitively expensive. It would cost at least $1950 in observer salaries alone to achieve a minimally acceptable level of generalizability (0.80). Travel, training, and other expenses would add substantial additional cost. Assuming that the Classroom Observation System could be made more reliable, especially by providing live observation training to observers, the cost of reaching minimally acceptable generalizability might only be cut in half.

- The Classroom Observation System may provide an especially sensitive measure of standards related to pupil engagement, but may not provide a complete measure of accomplished Standards-based teaching. An examination of the reasons observers gave to justify scores indicates, generally, that comments about "instruction" were weaker indicators of observers' scores than comments about "classroom environment." Positive comments about student engagement along with negative comments about classroom management explained more than a third of the variability in observers' scores (Adjusted $R^2 = .377$). Instruction in the areas of reading and discourse seem to be measured fairly well by classroom observation, but comments about writing and language study show low correlations with observers' scores.

- Comparisons with certification decisions based on exercise scores indicate that the exercises may not present a complete measure of accomplished Standards-based teaching and that classroom observation may provide important information that could contribute to the decision accuracy of the NBPTS certifications. Of the 15 candidates we observed, eight would be recommended for certification based on their exercise scores. Of those eight, four received poor or very poor scores on classroom observation. Generally, those four candidates were viewed by observers as having problems engaging students and managing their classes, two aspects of teaching that may not be well measured by existing EA/ELA exercises.

- The role of observation in NBPTS assessment merits further consideration.

EARLY ADOLESCENCE/ENGLISH LANGUAGE ARTS

CLASSROOM OBSERVATION STUDY

Robert Land and Corinne Harol
Center for the Study of Evaluation
UCLA

INTRODUCTION

During the 1993-94 school year, the first candidates for National Board for Professional Teaching Standards certification in Early Adolescence/English Language Arts (EA/ELA) constructed and submitted portfolios of nine exercises developed for the Board by the University of Pittsburgh Assessment Development Laboratory (ADL). The Classroom Observation Study was designed to develop and validate a Standards-based Classroom Observation System and to examine the costs and value-addedness of including an observation component in the mix of certification exercises.

This report presents detailed results regarding reliability of the Classroom Observation System; a general look at the costs associated with achieving various levels of generalizability using observation alone[1], data regarding the validity of classroom observation with respect to its grounding in the Standards and the perceptions of candidates; and a discussion of the degree to which classroom observation scores are consistent with certification decisions based on exercises recently scored by the Educational Testing Service (ETS).

The Classroom Observation Study was originally intended as a two-phase study, culminating in a study of a larger sample of current EA/ELA candidates and more days of observation. The Board's decision to eliminate Phase II funding necessitated truncating the planned study. Although carefully designed and implemented, this small Observation Study has several limitations beyond the obvious problems associated with the stability of statistics based on small samples. Teachers were not observed in their year of candidacy for certification, and some teachers had different assignments in

[1] Much of this validity and cost information was also presented in our Interim Report dated 13 April 1995.

the two years. Candidate and observer recruitment began only three weeks before the first live observations, and because of scheduling complexities too numerous to mention, most of the observations had to be completed in a three-week period in October 1994. As a result, training materials and observer training were severely abbreviated, and some observations had to be scheduled on dates that were less than ideal for the teachers.

Limitations notwithstanding, analyses of Observation Study scores presented in this report suggest that the Classroom Observation System provides a reliable vehicle for assessing teacher performance. Analyses presented in this report also suggest that scores obtained using the Classroom Observation System are adequately grounded in the Standards and that candidates perceive classroom observation as a highly valid form of assessment. Insofar as the reliability and validity of the Classroom Observation System scores can be assumed, comparisons with certification decisions based on exercise scores raise questions about whether either form of assessment presents a complete measure of accomplished teaching.

Description of the Study

The Design

Faced with limited resources, we sought advice regarding our research design at the August 1994 meeting of the Technical Analysis Group (TAG). We felt that we could observe only 15 or 16 teachers, but we were still interested in examining the reliability of observation with respect to the most likely sources of variability: teacher, observer, days of observation, and class ability level. Nearly identical designs were suggested by Robert Linn and Ross Traub. The designs called for six observers to visit, in pairs, 15 teachers over three days. Each of the 15 possible pairs of observers would work together three times, seeing two classes on a Day 1, a Day 2, and a Day 3 observation. Thus, each teacher would be seen twice by all six observers. Each observer would rate each class separately, so that each teacher would receive a total of 12 scores. Teachers were scored on a scale of 1 (low) to 4 (high), with plus and minus modifications permitted. To provide a degree of comparability, this scale was modeled on one adopted by ETS for scoring exercises.

Our design was implemented with several modifications made necessary by circumstances beyond our control. Because of a scheduling conflict, Day 1 and Day 2 observations were done out of order for one teacher. The last-minute withdrawal of one observer forced us to divide Observer 1's duties evenly among three observers who had been trained as backups. One of these backups was unable to complete her final observation and was replaced by an observer trainer for that day. This "composite" observer performed about the same as the full-time observers, with a mean over 30 observations of 2.55, just slightly higher than the overall mean of 2.46, and with an average discrepancy with partners' scores of 0.30, a bit under the overall average discrepancy of 0.37. For data analysis purposes, scores assigned by the substitute observers were treated as having been assigned by a single observer. Table 1 presents the Observation Study design as actually implemented.

Table 1

Design for Observation Study Data Collection

	Day 1	Day 1	Day 2	Day 2	Day 3	Day 3
	Segment	Segment	Segment	Segment	Segment	Segment
Teacher	1	2	1	2	1	2
1	1b,2	1b,2	3,4	3,4	5,6	5,6
2	1b,3	1b,3	2,5	2,5	4,6	4,6
3	1b,4	1b,4	2,6	2,6	3,5	3,5
4	1a,5	1a,5	2,4	2,4	3,6	3,6
5	1c,6	1c,6	2,3	2,3	4,5	4,5
6	3,4	3,4	5,6	5,6	1c ,2	1c,2
7	2,5	2,5	4,6	4,6	1c,3	1c,3
8	2,6	2,6	3,5	3,5	1d,4	1d,4
9	2,4	2,4	3,6	3,6	1c,5	1c,5
10	2,3	2,3	4,5	4,5	1b,6	1b,6
11	5,6	5,6	1a,2	1a,2	3,4	3,4
12	*1a,3*	*1a,3*	*4,6*	*4,6*	2,5	2,5
13	3,5	3,5	1c,4	1c,4	2,6	2,6
14	3,6	3,6	1b,5	1b,5	2,4	2,4
15	4,5	4,5	1a,6	1a,6	2,3	2,3

Note: Observer pairs are shown in the columns beneath the "Segment" headings. Italics indicate rater pairs who observed a particular teacher out of the planned order. The various raters who formed Composite Rater 1 are indicated by lowercase letters.

Candidate Sample

The Observation Study teachers were all 1993-94 candidates for EA/ELA certification. For logistical reasons, we observed only teachers in Southern California. In 1993-94, 28 Southern California teachers submitted completed portfolios for National Board Certification. Of those, seven were not teaching Early Adolescence/English Language Arts and one was no longer in the Southern California area in the fall of 1994. Two teachers did not wish to participate, and one teacher's schedule made participation unduly difficult. Because of scheduling concerns, we selected the final sample of 15 teachers based on geographic accessibility. The sample included 14 females—1 African American, 1 Asian American, 1 Latina, and 11 White—and 1 African American male. Fourteen taught in public schools, and 1 taught in a Catholic school. Nine taught in "urban" schools, and 6 taught in "suburban" schools. The 15 teachers in this sample reported an average of about 12 years of EA/ELA experience. Teaching assignments for this sample ranged from self-contained classrooms where the same students stayed with the same teacher all day to traditional periods where the teacher met with different classes at different levels throughout the day. Across the entire sample, classes of all ability levels with students from ages 10 to 15 were represented.

Teachers participating in the Observation Study agreed to be observed for three days for the same two class periods each day. Teachers with classes of differing ability or grade level agreed to allow visits to their higher and lower level classes. The three teachers who met with only one group of students agreed to be observed over two distinct instructional segments of at least 40 minutes each, each day. Teachers were asked to demonstrate a range of content over the three days, focusing on a post-reading (Post-Reading Interpretive Discussion Exercise or PRIDE-like) lesson and a writing lesson at least once each.

Preliminary analyses of Observation Record Sheets show that all teachers at least touched on each of the content areas identified in the EA/ELA Standards (reading, writing, discourse, and language study), but that reading instruction dominated. Teachers integrated lesson-types into their curriculum as made sense for their particular classes; consequently, lesson-types were not systematically distributed across days. Teachers agreed to complete Teacher

Demographics and Class Demographics Information Forms and to respond, at least briefly, to pre- and post-teaching interview questions on the Observation Cover Sheet. Teachers also agreed to participate in a follow-up telephone interview. The teachers understood and were frequently reminded that the classroom observations were for research purposes only and would have no bearing on whether they would be certified. All of the teachers who participated were cooperative and seemed to welcome having observers in their classes. Their main complaint was that they would not be able to receive feedback from the observers.

Observers

Originally, we hoped to use the most consistent observers of those who had worked with pilot versions of the Observation System, but all of these observers had obligations that prevented them from observing full-time during the fall of 1994. In addition, members of the National Advisory Committee, who met in early September, strongly recommended that observers be teachers who were accomplished, practicing EA/ELA teachers themselves. To recruit such observers, we consulted the Center's data base and contacted a number of outside sources including graduate schools of education, professional organizations, district supervisors, and literature and writing project directors. We were able to identify a number of suitable candidates; most of them were very interested but were already overcommitted or unwilling or unable to take three weeks away from teaching, at least on short notice. In the end, we were able to recruit, retain, and train eight observers, who averaged just under 11 years English language arts teaching experience. Five observers were active EA/ELA teachers, recommended by supervisors as highly accomplished. Two observers were former EA/ELA teachers, one still active as a curriculum consultant and the other as a Grade 5 to 8 substitute teacher and volunteer teaching and directing a middle-school “great books” program. The remaining observer was a graduate student with extensive large-scale research experience, several years college-level teaching experience, and more than a year's experience as a middle-school English tutor. We strove for diversity in our selection of observers: two were Latino/a, one was Asian American, and one was African American. Their ages ranged from mid-20s to early 60s, and their teaching experience included urban, suburban, and Catholic schools. Seven of the eight observers were women, a proportion similar to that of the EA/ELA teaching population.

Observer Training

Observers were trained using a selection of Post-Reading Interpretative Discussion (PRIDE) and Planning and Teaching Exercise (PTE) videotapes of teachers from those judged by ETS and the ADL to exemplify a range of performances on a 1-to-4 scale. Training consisted of:

1. an overview of the certification process, the EA/ELA Standards, and the role of the Observation Study;

2. an introduction to the observer's task and a review of the Observation Instrument;

3. a review of sample, completed instruments;

4. a trial use of the Observation Instrument with a video segment, followed by a discussion of sample instruments completed for that segment;

5. a series of uses of the Observation Instrument with full videos by individual trainees, followed by pair and whole group consensus-building on both salient features and scores.

Four observers who had used the instrument during the August 1994 piloting were present during the training. They provided experience-based explanations of the use of the instrument, discussed common problems confronted during observation, and coached trainee pairs as they worked toward consensus.

The centerpiece of the training was the "Standards Capturing Framework." This framework emerged as our response to the challenge of helping observers avoid "checklist" scoring while remaining grounded in the Standards. The framework was based, in part, on the reports of experienced observers and on the responses of experienced observers and teachers to a survey asking them to assess the observability and importance of roughly 350 behaviors noted in the EA/ELA Standards.

Observer trainees were directed to use the framework for two purposes:

1. to avoid fixing their attention too long on any particular aspect of teaching by regularly shifting their focus from Environment to Instruction, and

2. to increase their use of Standards-like language as they recorded events by keeping in mind the precepts that seemed especially in keeping with the spirit of the Standards: Appropriateness, Inclusiveness, and Connectedness.

Observer trainees were also directed to place their principal focus on recording what they saw, not on coding to the Standards. Coding events to the Standards served as a check on their Standards-based orientation.

Observer trainees were asked to maintain high standards. They were told that scores below 3 did not necessarily mean that the observed teacher was “bad” or even uncertifiable. They were directed to give a score of 1 (“weak”) if the performance, as a whole, either did not reflect National Board Standards-based teaching (regardless of how “good” it might be given other standards), or reflected inept, inappropriate, or ineffective attempts at Standards-based teaching. They were directed to give a score of 2 (“adequate”) if the performance, as a whole, reflected uneven Standards-based teaching (even if the overall impression of the class was generally satisfactory). They were directed to give a score of 3 (“accomplished”) if the performance, as a whole, reflected consistent Standards-based teaching (even if there were some occasional weaknesses). They were directed to give a score of 4 (“exemplary”) if the performance, as a whole, was consistently and convincingly Standards-based.

Training lasted six hours and took place on a Friday; live observations began the following Monday. Twice during the live observations (after the first day, and at the halfway point), trainers met with observers to discuss general problems, taking care not to mention particular teachers by name. Trainers were also available throughout the data collection period to provide support to individual observees. Observation Instruments were collected and examined as completed to identify potentially serious problems. (No substantive changes to the observation procedures were made as a result of these contacts.)

The Score Scale and Scoring Rules

Each class or instructional segment was scored independently by two observers. Scoring took place as soon as possible after the end of each class or instructional segment. Using the Observation Decision Sheet, observers first assigned a "band score" of 1/2, 2/3, or 3/4; then they reviewed their notes to identify and record the salient features of the teacher's performance that led them to that first impression. Finally, observers settled on a final score based on their more systematic reflection about the teacher's performance. A candidate's final score on each class segment was rated on a scale of 1 to 4. Minus (-) and plus (+) scores were permitted, but observers were encouraged to use ordinal numbers. In theory, observers could have used a 12-point scale ranging from -1 to 4+, but in practice they used a 10-point scale ranging from a minimum of 1 to a maximum of 4. For data analysis purposes, the use of a minus diminished the ordinal score by 0.25, and the use of a plus increased the ordinal score by 0.25.

Basic Statistics Describing Scores

In all, 180 separate scores were collected. Tables 2, 3, and 4 describe these scores in terms of central tendency and dispersion.

It is apparent from these tables that assigned scores covered the range of possible scores (with ordinal numbers making up the majority of scores), indicating that assessors were able to discriminate between levels of performance. They also show that score means did not vary much by class segment within day.

Table 2

Score Means by Observation Day

	Mean	Std. Dev.	Std. Error	Minimum	Maximum	Median
Day 1 (*n*=60)	2.471	0.915	0.118	1.000	4.000	2.000
Day 2 (*n*=60)	2.733	0.933	0.120	1.000	4.000	3.000
Day 3 (*n*=60)	2.171	0.851	0.110	1.000	4.000	2.000

Table 3

Score Means by Segment Within Observation Day

	Mean	Std. Dev.	Std. Error	Minimum	Maximum	Median
Day 1 Segment 1 (N=30)	2.475	0.925	0.169	1.000	4.000	2.000
Day 1 Segment 2 (N=30)	2.467	0.921	0.168	1.000	4.000	2.375
Day 2 Segment 1 (N=30)	2.825	0.889	0.162	1.000	4.000	3.000
Day 2 Segment 2 (N=30)	2.642	0.982	0.179	1.000	4.000	2.875
Day 3 Segment 1 (N=30)	2.100	0.747	0.136	1.000	3.250	2.000
Day 3 Segment 2 (N=30)	2.242	0.950	0.173	1.000	4.000	2.500

Table 4

Frequency Distribution of Scores

Score	Count
1.000	25
1.250	4
1.750	4
2.000	59
2.250	4
2.750	5
3.000	44
3.250	9
3.750	2
4.000	24
Total	180

In the original design it was supposed that observers would be watching two very different classes for each teacher, and that scores for these classes might vary according to the teacher's skill in teaching each level. We were able to observe only five teachers who had classes of markedly differing ability between segments. Scores for all five of these teachers did differ for the same observer between segments, with an average difference of 0.66. Scores differed for the same observer between segments for eight of the remaining 10 teachers, with an average difference 0.26. Admittedly, these data are based on an extremely small sample; however, they do suggest that examining teachers' performances over student groups with markedly different levels of ability might be important to obtaining an accurate picture of their teaching ability.

Scores did vary by observation day, with score means from Day 2 higher than Day 1 means and significantly higher (Tukey's HSD contrast $p < 0.05$) than Day 3 means. Of the several plausible explanations for score differences by Day, an especially intriguing one is the "TGIF" effect. Because of time constraints on completing the observations before holidays, a disproportionate number of Day 2 observations happened to be scheduled on Fridays, when somewhat looser classroom structure and more planned student interaction may have led to higher scores. Another possible explanation is that teachers may have disrupted their routines to demonstrate lesson-types on Day 3 that they had been unable to show over the first two days of teaching. These disruptions may have negatively affected student attitude or may have stretched the teachers' skill at integrating curriculum too far. A third possibility is that some teachers ran out of "showcase" lessons after two days, and that Day 3 reflects their routine performance level.

Inter-Observation Correlations

Table 5 shows that low correlations exist between candidates' Day 1 and Day 2 or 3 scores, but that Days 2 and 3 are moderately correlated. Low correlations are not surprising, given that different teachers taught different types of lessons each day. "Reading Discussion" lessons comprised 46 of the 90 lessons observed, and these "PRIDE-like" lessons received significantly higher scores than did other lesson types (2.71 vs. 2.20, $p < .01$, $df_{1,150}$). ANOVA results also show a significant Teacher x Lesson-Type interaction ($p < .001$, $df_{14,150}$).

Table 5

Correlations Among Teachers' Average Daily Scores

Average score	Day 1	Day 2	Day 3
Day 1 (*n*=15)	1.000		
Day 2 (*n*=15)	0.172	1.000	
Day 3 (*n*=15)	0.173	0.509	1.000

Table 6 shows that mean segment scores were highly correlated on the same day (in bold face), but were not as highly correlated across different days. This pattern of correlations may result partly from the fact that the same pair of observers scored both segments for a single teacher on a given day. Also, as noted above, the two segments were quite similar for many of the teachers, and many of these teachers taught the same lesson for both segments on a given day but taught very different lessons across the three days.

Table 6

Correlation Among Teachers' Average Segment Scores for Each Day

		Day 1		Day 2		Day 3	
Day	Segment	Seg. 1	Seg. 2	Seg. 1	Seg. 2	Seg. 1	Seg. 2
Day 1	Seg. 1 (*n*=15)	1.000					
Day 1	Seg. 2 (*n*=15)	**0.784**	1.000				
Day 2	Seg. 1 (*n*=15)	0.241	0.124	1.000			
Day 2	Seg. 2 (*n*=15)	0.099	0.096	**0.805**	1.000		
Day 3	Seg. 1 (*n*=15)	0.303	0.077	0.482	0.391	1.000	
Day 3	Seg. 2 (*n*=15)	0.228	0.022	0.447	0.415	**0.723**	1.000

Like Table 5, Table 6 shows higher correlations among Day 2 and Day 3 scores. Given that score averages for Days 2 and 3 were significantly different, and given that most teachers taught different lesson types on Days 2 and 3, these higher correlations suggest that raters may have become more reliable as they gained experience or that teachers became more consistent as a result of increased comfort with observers in their classes by Day 2.

Some Statistics Indicative of Reliability

Following are the results of several analyses that suggest that the EA/ELA Classroom Observation System provides acceptable levels of interrater reliability.

Distributions of Inter-Assessor Score Differences

Table 7 contains frequency distributions, proportion frequency distributions, and cumulative proportion frequency distributions of differences between the scores assigned by two assessors to the class segments they observed. They suggest the degree to which different assessors, working independently, interpreted candidates' performances similarly.

Table 7

Distribution of Differences Between Raters' Scores

Differenc e between scores	Count	Percent	Cumulative proportion
0.00	55	61.11	0.61
0.25	7	7.78	0.69
0.50	2	2.22	0.71
0.75	8	8.89	0.80
1.00	13	14.44	0.94
1.50	1	1.11	0.96
2.00	3	3.33	0.99
2.75	1	1.11	1.00
Total	90	100.00	

Tables 8 and 9 show levels of agreement across days and within segments. These tables show a moderate level of exact agreement, and they suggest that levels of agreement were fairly steady over observation days.

Table 8

Cumulative Distribution of Inter-Assessor Agreement by Day and Segment

	% Identical	% Within 0.25 point	% Within 0.5 point	% Within 1.00 point
Overall	61	69	71	94
Day 1 Segment 1	87	87	87	93
Day 1 Segment 2	60	60	60	86
Day 2 Segment 1	40	60	60	93
Day 2 Segment 2	60	73	73	93
Day 3 Segment 1	60	73	73	100
Day 3 Segment 2	60	60	73	100

Table 9

Inter-Assessor Agreement by Day and Segment: Assigned Scores Converted to Ordinal Scores

	% Same ordinal score	% 1-point differences	% 2-point differences
Overall	70	94	99
Day 1 Segment 1	87	93	93
Day 1 Segment 2	60	87	100
Day 2 Segment 1	60	93	100
Day 2 Segment 2	73	93	100
Day 3 Segment 1	73	100	100
Day 3 Segment 2	67	100	100

Generalizability of Scores Across Assessors

The following tables show a fair level of reliability by both day and class. Generally increasing levels of reliability may suggest that observers improved with practice and that observers may have needed more training before doing their first "real" observations. Another possible explanation is that observers used less of the range of scores Day 3, Segment 1, thereby creating fewer outliers.

Table 10

Variance by Observation Day

	Variance component			Generalizability coefficient	
Day	Candidate	Assessor	Error	One rater	Two raters
1	0.502	0.019	0.251	0.65	0.79
2	0.609	0.006	0.196	0.75	0.86
3	0.570	0.000	0.072	0.89	0.94

Table 11

Variance by Observation Day and Class or Segment

		Variance component			Generalizability coefficient	
Day	Seg.	Candidate	Assessor	Error	One rater	Two raters
1	1	0.561	0.053	0.245	0.65	0.79
1	2	0.524	0.001	0.341	0.61	0.75
2	1	0.532	0.023	0.254	0.66	0.79
2	2	0.798	0.000	0.194	0.80	0.89
3	1	0.469	0.005	0.103	0.81	0.90
3	2	0.823	0.010	0.122	0.86	0.93

Inter-Assessor Correlations

Correlations between scores assigned to the 15 candidates by the pairs of assessors within segments and within days show levels of reliability similar to those indicated by generalizability results. Table 12 contains estimates of inter-assessor correlations. They suggest quite reasonable degrees of agreement between assessors who reviewed the same performances of the same candidates. Increases in correlations over days again suggest that more training might have improved initial observer reliability.

Table 12

Correlations Between Raters' Scores by Segment and Day

Day	Segment	Correlation
1	1	0.691
1	2	0.611
1	Both	0.619
2	1	0.641
2	2	0.797
2	Both	0.725
3	1	0.817
3	2	0.866
3	Both	0.828
All	Both	0.737

Generalizability and Preliminary Cost Estimates

Tables 13 and 14 present generalizability projections for various combinations of raters and days based on observation data. Results presented in Table 13 are based on individual observers' average scores for each teacher on each observation day. That is, segment scores were averaged within observer and day, yielding 90 mean scores for all 15 teachers. Table 14 removes the Day 1 scores, treating Day 1 observations as additional training and practice.

Table 13

G-Study Variance Components, D-Study Generalizability Levels, and Salary Costs: All Days

No. of raters	No. of days	Absolute generalizability	Relative generalizability	Salary costs
1	1	0.24	0.25	$150
1	5	0.61	0.63	$750
1	9	0.73	0.75	$1350
1	13	0.80	0.81	$1950
2	1	0.31	0.34	$300
2	2	0.47	0.50	$600
2	3	0.57	0.60	$900
2	4	0.64	0.67	$1200
2	6	0.73	0.75	$1800
2	7	0.76	0.78	$2100
2	8	0.78	0.80	$2400
2	9	0.80	0.82	$2700
2	10	0.82	0.83	$3000
3	1	0.35	0.38	$450
3	3	0.61	0.65	$1350
3	5	0.72	0.75	$2250

Source:	
Teacher	0.1862
Day	0.0481
Teacher x Day	0.1795
Rater : (Teacher x Day)	0.3774

Throughout this report, it is suggested that observers may have been undertrained. Setting aside any number of very plausible alternative explanations, results presented in Table 14 again suggest that additional rater training may have improved the initial reliability of observation, perhaps substantially. Both tables show that number of days, not number of observers, is the biggest factor in increasing generalizability.

Table 14

G-Study Variance Components, D-Study Generalizability Levels, and Salary Costs: Days 2 and 3 Only

No. of raters	No. of days	Absolute generaliz-ability	Relative generaliz-ability	Salary costs
1	1	0.39	0.46	$150
1	2	0.56	0.63	$300
1	5	0.76	0.81	$750
1	6	0.79	0.84	$900
1	7	0.82	0.86	$1050
2	1	0.42	0.51	$300
2	2	0.59	0.67	$600
2	5	0.78	0.84	$1500
2	6	0.81	0.86	$1800
2	7	0.84	0.88	$2100
3	1	0.43	0.53	$450
3	2	0.60	0.69	$900
3	5	0.79	0.85	$2250

Source:

Teacher	0.3340
Day	0.1366
Teacher x Day	0.2556
Rater : (Teacher x Day)	0

In both tables, generalizability projections have been used to estimate costs. Cost estimates assume that raters will be paid $150 per day and that only one teacher can be observed on any given day. Each day is assumed to include two classes or instructional segments scored separately by the same observer(s) with daily scores averaged, as they were for Observation Study data. Under rare circumstances, it might be possible for an observer to see more than one teacher in a day, thus reducing costs. Other options, such as making classroom observation part of a district's or state's professional development program, might reduce costs as well. Even so, the costs listed below would almost certainly be much higher when incidental expenses for local travel, materials duplication, telephone contacts, and the like are added to more substantial supervision, support, and management costs. Moreover, substantial travel costs would very likely be associated with at least some observations.

Training eight local observers cost us about $1800 in observer and trainer salaries, but our training session was probably too short, and four of our trainers worked for lunch and a half-day's salary. Increasing training would very likely increase the reliability of observation and might, therefore, reduce overall cost. Assuming three days of training at $150 per day for trainers and observers, and assuming one trainer for eight observers, and assuming that no observer drops out of training or fails to qualify, training one observer would cost about $500 in salaries alone. Training of a nationally representative sample of observers would involve substantially increased costs, just for travel.

As Table 13 shows, it would cost roughly $2000 in observer salaries (one observer for 13 days) to reach a minimally acceptable level of generalizability for a high-stakes decision. If the higher levels of reliability suggested by Table 14 could be achieved, observer salaries would cost less than half as much, about $900, to reach the same level of generalizability. Of course, no one envisions classroom observation as the only assessment for National Board Certification. And indeed, our small-scale experience with scheduling observations suggests that the logistic difficulties of scheduling more than three or four visits may prove overwhelming, especially on a large scale. In our experience, the odds of being able to schedule a classroom visit on any given day were about two in three—this during a three-week period uninterrupted by holidays.

Validity Results

Severely reduced funding and a compressed project schedule led to the elimination of several validation efforts originally proposed. This section presents selected results of candidate interviews; a description of the relationship between "salient features" and the Standards; analysis of the correspondence between lists of "salient features" for observer pairs who saw the same classes; an analysis of the degree to which observation scores are consistent with certification decisions based on exercise scores; and a hypothesis for the discrepancy between the exercise scores and observation scores.

Teachers' Impressions of the Validity of Observation

After completing the observations, we were able to interview 14 of the 15 teachers. Overwhelmingly, teachers feel that observation is a valid measure of their teaching, with most of them clearly indicating that, in their opinion, observation is more valid than the EA/ELA exercises they submitted.

All of the teachers said that classroom observation provided an opportunity to demonstrate important aspects of their teaching to a “moderate” or to a “substantial” extent. We asked the teachers what they felt could be achieved in a classroom observation that was not accomplished by the portfolio. More than any other comment, teachers said that student-teacher interaction can only be captured in a live observation. One teacher said, “You can have all the theories, but if you can't make it meaningful to the students, then it is useless.” Another said that you can see how the teacher “handles the class in difficult or spontaneous situations.” Another said that observation evaluates “the real product of teaching. . . . It's not an abstraction. If you can't get along with kids, you can't do it.” One teacher said that an observer in the classroom can see “a glance or an interaction” between a teacher and a student, and compared this with a video, which cannot capture the teacher and the students at the same time.

Many other teachers also compared the video exercises with the classroom observation. In addition to many comments about the superiority of live observation in assessing student-teacher interaction, several teachers felt that live observation is more fair because it is not subject to distortion by unequal access to technical support; one teacher said, “Live observation is a lot more valid than video. A video can be set up and manipulated, and thus may not present a typical day.” Many also said that live observation was much less worry and work on their part than preparing a video, and one teacher commented that a video camera is more disruptive to the classroom than a live observer. Two teachers said the observation is better than the video because the observer can get a feel for the school environment as a whole.

One teacher said she was frustrated by the “letter of the law” of the portfolio exercises. She felt that she had more control over what was presented in the observation and that the observers could get more of a "multi-dimensional, sensory image of the teacher," compared with the "flat" portfolio exercises.

When asked about the limitations of classroom observation, most teachers worried that the context, continuity, and progress of the class could be lost in isolated observations. As one teacher remarked, "Observation cannot show the way that a teacher constructs instruction over the long term in order to meet the particular needs of a classroom." Many teachers felt that the observations should be stretched out over the course of a year because they felt that it would be important for the evaluators to be able to see how the class progresses. One teacher said, "At the beginning of the year, the teacher is in the process of creating an environment. Coming over the course of a year, an observer could see how the teacher sets up the environment, how the environment contributes to learning, and whether the students have improved over the year." Another teacher said that the teacher and the students must together establish a "safe zone" where students feel confident and which produces the "teachable moments that define a class." This teacher felt that observations over the course of a year would provide more opportunity to see this.

Some teachers worried that observation would not allow them to present their philosophy of education, and one said she would like a chance to show what her idea behind a lesson was, even if the execution turned out poorly. Several teachers felt that examples of student work as well as student evaluations or interviews should be part of any evaluation of teachers, and one teacher would want evaluators to know about a teacher's involvement in the community.

Many teachers were skeptical of the validity of the school site and assessment center exercises. Referring to the assessment center exercises, one teacher said, "They were interesting, but had nothing to do with teaching." Nonetheless, many teachers also commented that completing the exercises was a valuable learning experience.

Salient Features and the EA/ELA Standards

Observers were told that ideally they would give the same scores for the same reasons (called "salient features" on the Observation Decision Sheet) and that those reasons could be traced to the language of the EA/ELA Standards.

Observers used the “Standards-based Reference List” of 105 descriptors of accomplished teaching drawn from the EA/ELA Standards as a common resource. Otherwise, they were free to list as many salient features as they thought were necessary to justify their score, and they were free to cite bases for their score at the level of individual descriptors or at the level of the 10 subcategories (listed on the “Standards Capturing Framework” under Environment and Instruction) under which individual descriptors were subsumed. They were also free to note something as a “salient feature” whether or not they could find it on the lists, so long as they felt that the feature was in keeping with the Standards.

The annotated “Standards-based Reference List” in the appendixes of this report shows the location of each descriptor in the Standards and indicates the number of times a particular descriptor was cited as a salient feature by the observers. In all, 1444 salient features were cited on 180 separately completed decision sheets. Experienced observers reviewed all salient features and verified their classification and further classified them as “positive,” “negative,” or “mixed.” Of the 1444 salient features listed, 1019 were classified as “positive,” 379 as “negative,” and 46 as “mixed.” Twelve could not be classified. Of the 105 descriptors available, 95 were cited at least once. All subcategories were cited, with “Language Study” receiving the fewest (46) and “Engagement” receiving the most (279) citations.

Figure 1 reports the total number of times each subcategory was cited (see Appendix D for more complete descriptions).

Ignoring “mixed” comments and multiple cites within a single subcategory, observers listed an average of 5.6 out of 20 possible subcategories (10 positive and 10 negative) as salient features. Comparisons of salient features listed by observers who scored the same segment independently show an average of 2.85 agreements, more than might be expected by chance alone. In the 63 out of 90 cases where observers agreed on the same ordinal score, observers agreed on 3 of 5.6 cites on average.

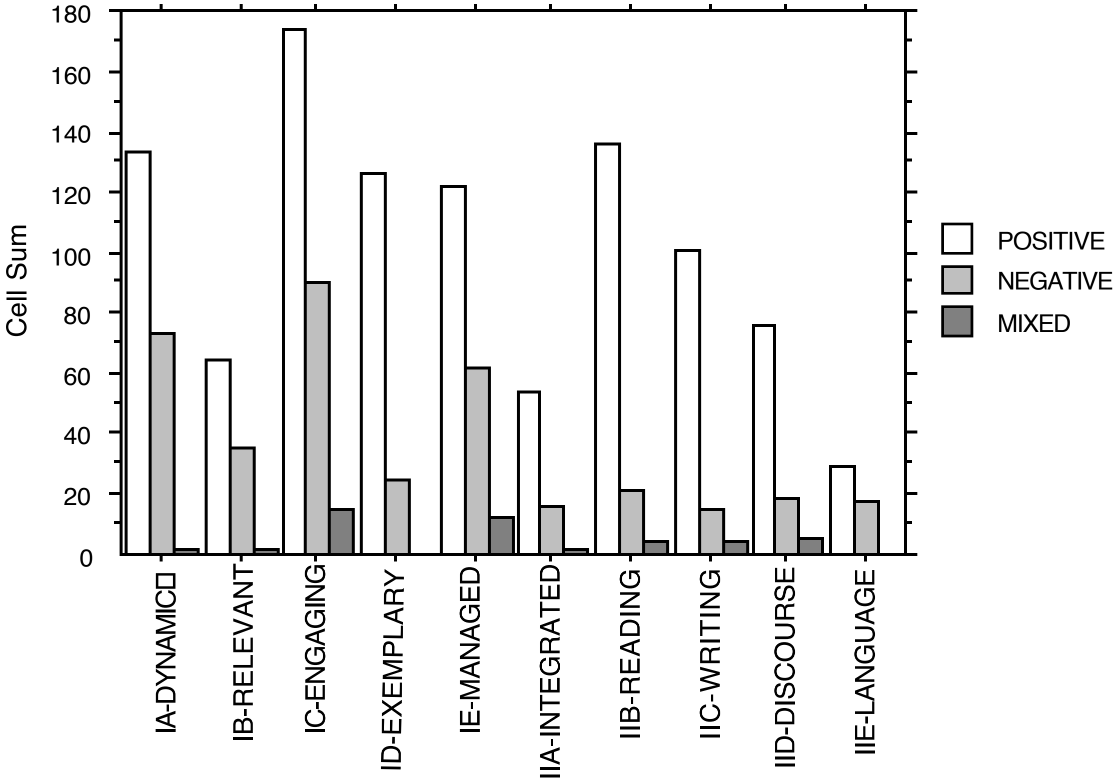


Figure 1. Total number of times each subcategory was cited.

Tables 15 and 16 present the correlations between final Observation Study scores and the ten major salient feature subcategories, broken down by positive and negative comments.

Table 15

Correlations Between Salient Features Cited and Average Observation Score: Positive Comments

	IA+	IB+	IC+	ID+	IE+	IIA+	IIB+	IIC+	IID+	IIE+
Score	.408	.249	.468	.285	.411	.375	.414	.167	.357	.036

Table 16

Correlations Between Salient Features Cited and Average Observation Score: Negative Comments

	IA-	IB-	IC-	ID-	IE-	IIA-	IIB-	IIC-	IID-	IIE-
Score	-.397	-.285	-.386	-.382	-.505	-.176	-.189	-.216	-.191	-.121

These tables show that negative comments on IE (Classroom Management) and positive comments on IC (Student Engagement), IIB (Reading), and IE (Classroom Management) were most highly correlated with overall score. The two most highly correlated subcategories (negative comments on Classroom Management and positive comments on Student Engagement) explain about 38% of the variability in score (Adjusted R^2 = .377).

Comparison of Certification Decisions and Observation Results

Comparison of Observation Score Patterns With Exercise Scores. One simple and revealing way to compare exercise scores and Observation Study scores was suggested by Linda Crocker at the May 1995 meeting of the TAG. She suggested presenting a table that showed the number of times candidates received observation scores of 1, indicating weak teaching performance, next to a column showing the certification recommendation of the Standard Setting Committee. Table 17 shows, in order from low to high scores on the exercises, the candidates' final scores and certification decisions along with the distribution of the ordinal scores they received in the Observation Study.

Table 17 also shows substantial discrepancy between the exercise scores and the Observation Study scores. Of the 29 observation scores of 1 (weak), 19 were given to candidates who were recommended for certification. The third highest ranking candidate received eight of the 19 “ones.” Twelve of the 26 scores of “four” (exemplary) were given to candidates not recommended for certification.

Table 17

Frequency of Ordinal Observation Scores Compared With Ranked Exercise Scores

ID	Average weighted exercise score	Certification recommendation	Frequency of Observation Study ordinal scores				Hypothetica lobservation certification recommendation
			Ones (Weak)	Twos (Adequate)	Threes (Accomplished)	Fours (Exemplary)	
J	No score	No	0	0	6	6	Yes
K	No score	No	0	10	2	0	No
C	184.54	No	4	8	0	0	No
M	203.46	No	4	3	4	1	No
A	208.34	No	0	4	6	2	Yes
B	253.09	No	0	6	5	1	Yes
E	266.15	No	2	5	3	2	No
N	276.29	Yes	4	2	4	2	No
H	286.45	Yes	4	5	3	0	No
L	296.89	Yes	0	6	2	4	Yes
G	299.30	Yes	0	5	3	4	Yes
I	301.18	Yes	0	3	9	0	Yes
D	302.00	Yes	8	4	0	0	No
O	321.48	Yes	3	4	5	0	No
F	348.94	Yes	0	2	6	4	Yes

Comparison of Pass/No Pass Decisions

The overall mean of Observation Study scores was 2.48. We decided on a cut score for passing of 2.6; this seems a logical cut point because the scores clustered into four groups, and 2.6 separated the two higher and two lower clusters. Moreover, 2.6 is both higher than the observed average and indicative of teaching that was better than a theoretical average of 2.5. Use of a higher cut score was rejected as overly stringent, compared with the cut score set for exercises. Using the cut score of 2.6, 7 out of 15 candidates would, hypothetically, be certified based on observation scores alone. On a scale of 100 to 400, candidates with a weighted average of exercise scores 275 or above were recommended for certification by the Standards Setting Committee. In this sample, the observed mean weighted score was 273, only slightly lower than the standard for passing. Eight of the 15 candidates were recommended for certification. In only four of those eight cases did hypothetical Observation Study certification decisions agree with the certification recommendations based on exercise scoring. Table 18 shows a matrix of agreements on the two sets of scores.

Table 18

Comparison of Candidate Certification Recommendations With Hypothetical Observation Score Certification Recommendations

	Observation		
Exercises	Pass	Not pass	Total
Pass	4	4	8
Not pass	3	4	7
Total	7	8	15

Table 18 reports on decisions for all 15 candidates, two of whom were not recommended for certification because of incomplete portfolios. On the surface, the table shows inconsistencies between certification decisions based on exercises and hypothetical decisions based on observation. We analyzed the performances and situations of the candidates whose scores and certification recommendations were discrepant in order to discover what could have led to so many disagreements.

Through observation, we might have certified three teachers who were not recommended for certification based on their portfolio scores. An analysis of the cases of these three teachers suggests that “false negatives” (failing to certify someone who is an accomplished teacher) does not seem to be a serious problem with the exercise scoring at this time.

- Candidate B received an average observation score of 2.625, a borderline pass in our estimation The observers commented positively on her relationship with her students, something we believe is better measured by observation than by the exercises. Most of the negative comments for this teacher had to do with instruction that was either inappropriate for the students or that was not Standards-based, something that may be weighted more heavily in the portfolio exercises. We believe a composite of the two scores would accurately reflect her performance, and thus would agree with the exercise-based certification recommendation.

- Candidate A is a relatively inexperienced teacher (five years experience at the time of the observations), who described the portfolio process as formative and who described herself as a much better teacher after going through the certification process and getting one more year of experience. Also, she received very positive comments from the observers on classroom environment and on her relationship with her students, things that, again, we believe are better measured in observation. We believe that this candidate would pass the certification process in the future, especially if classroom observation (or another clear measure of student engagement) were included. This candidate may have been someone whose maturation from one year to the next explains the difference in her performance.

- Candidate J was the highest scoring candidate in the Observation Study, but her PRIDE video was considered unscorable. Had this candidate submitted a scorable PRIDE exercise and had she not passed, her case would point to a more serious question about “false negatives.”

In our opinion, “false negatives” are not as great a threat to the credibility of the certification process as are “false positives” (certifying someone who is not accomplished). An analysis of the candidates who passed the exercise scoring but did not score as well on observation indicates that false positives may be a concern.

Four candidates who were recommended for certification by the EA/ELA standard setting committee probably would not have been recommended for certification based on classroom observations alone. We can think of plausible explanations for the discrepancy in three cases:

- Candidate H had a new assignment for the school year of the observation; she described her assignment as punitive, resulting from problems with school administrators. The students for her new assignment had been transferred from other schools for discipline problems. This teacher’s scores were highly variable, and many observers cited their observation of this teacher as the most difficult observation to score.

- Candidates N and O had similar profiles. Both had a wide variety of scores and both had one day in particular that dramatically lowered their average score. In both cases, we were not able to observe two different groups of students (one candidate has a self-contained classroom and for the other candidate, we only had access to one EA/ELA class, a double period). The impact of a bad day (whether the result of the teacher, students, or lesson) would be significantly magnified. Hence, these teachers did not have the same opportunity as others to demonstrate successful teaching with different groups of students. The relatively low scores on the classroom observations might then be a result of one "bad day." Should classroom observation be used, we believe this problem would have to be addressed.

- One candidate, however, appears to have received a positive certification score that would be disconfirmed by direct observation using the Classroom Observation System. Unlike Candidates H, N, and O, the differences in scores for this candidate, D, cannot be explained by differences in assignment, students, or instructional practice (indeed, one of the lessons we observed for Candidate D was identical to the PTE lesson she submitted). Moreover, her score pattern was consistent: Of the six observers, none awarded this candidate a passing score. Eight out of 12 times the class was deemed "weak" and 4 times it was scored as "adequate." The six observers gave highly consistent reasons for their consistently low scores. She was the lowest scoring candidate (out of 15) for the Observation Study and yet was the third highest scoring candidate (out of 13) for the portfolio exercises. If there were a pattern of false positives, one would expect to find, perhaps even in a sample this small, a convincing "false positive." We believe that this case (Candidate D) is a convincing and troubling "false positive."

Salient Features and Pass/No Pass Decisions

As discussed above, certain features of a teacher's performance were especially influential in determining the observation score. In particular, negative comments on Classroom Management and positive comments on Student Engagement most strongly influenced observation scores. We call these two areas measurements of student/teacher interaction, and we hypothesized that this could most effectively be measured in observation. In order to determine whether the discrepancies between observation scores and exercise scores were systematic or random, we analyzed the four groups of teachers from the last section according to the observers' comments on these "salient features." Tables 19 and 20 report the results of these analyses.

Table 19

Average Incidence (Per Teacher) of Negative Comments on Classroom Management

	Observation		
Exercises	Pass	Not pass	Total
Pass	1.25 (*n*=4)	7.25 (*n*=4)	4.25 (*n*=8)
Not pass	0.67 (*n*=3)	6.50 (*n*=4)	4.01 (*n*=7)
Total	1.00 (*n*=7)	6.88 (*n*=8)	4.13 (*N*=15)

Table 20

Average Incidence (Per Teacher) of Positive Comments on Student Engagement

	Observation		
Exercises	Pass	Not pass	Total
Pass	14.26 (*n*=4)	10.00 (*n*=4)	12.12 (*n*=8)
Not pass	16.67 (*n*=3)	6.74 (*n*=4)	11.00 (*n*=7)
Total	15.29 (*n*=7)	8.38 (*n*=8)	11.60 (*N*=15)

The relationship is more evident in Figures 2 and 3:

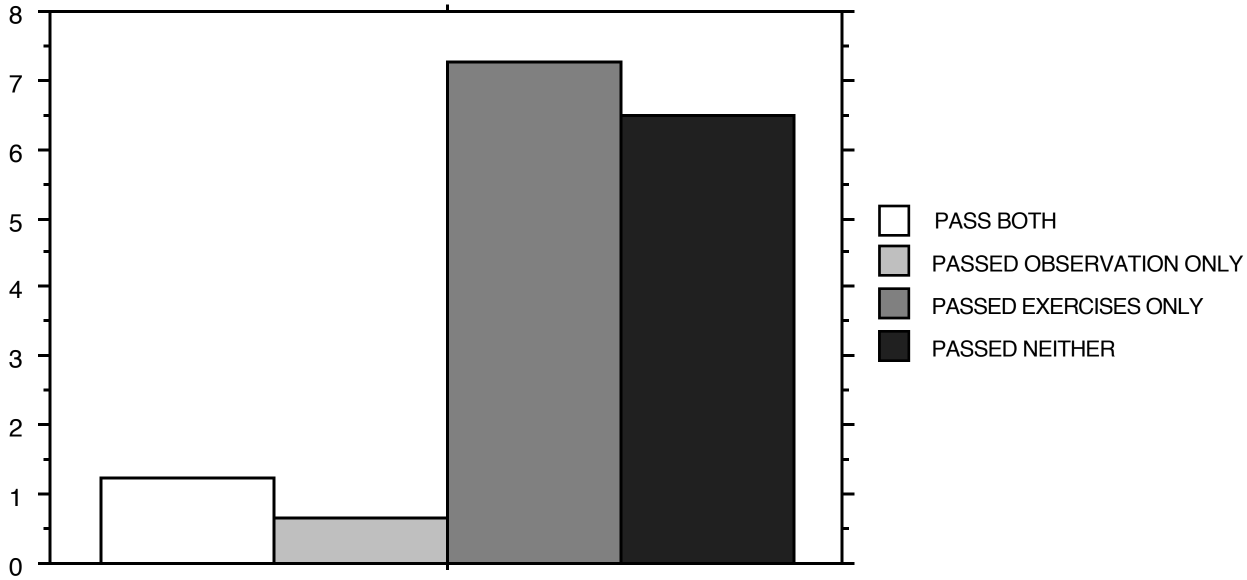


Figure 2. Incidence of negative comments on classroom management.

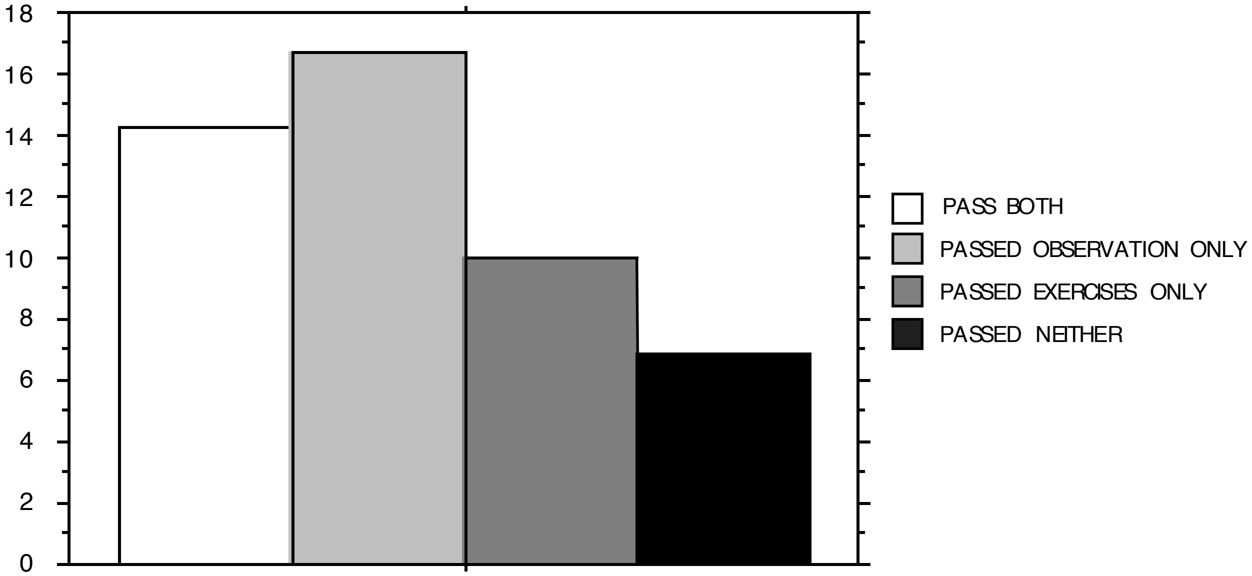


Figure 3. Incidence of positive comments on student engagement.

Table 19 shows that the group with highest average number of positive comments on Student Engagement is that group of four teachers who scored high on the observation but were not recommended for certification based on their exercise scores. Table 20 shows that the group of teachers who scored high on the observation but not on the exercises had very few negative comments on Classroom Management, even fewer than the group that scored high on both measures. Moreover, the group with the largest average number of negative comments on Classroom Management is the group that scored well on the exercises but not on the observation. These tables would tend to confirm a hypothesis that the discrepancy between the Observation Study scores and the exercise scores may be the result of differences in the way the two forms of assessment measure and value student/teacher interaction.

Exercise and observation scores may each provide a partial view of accomplished teaching. Indeed, significant elements of accomplished teaching may not be captured by either measure. (In particular, the teaching of writing may be inadequately assessed, considering that the Analysis of Student Writing exercise was not scored and that no writing lesson received an observation score of 4.)

Comparison of Exercise Videos and Observation Study Scores

To further explore the cases where observation results seem inconsistent with certification decisions, we had an observer, who had seen all of the candidates in the classroom, review PRIDE and/or PTE videos submitted by the candidates in question in order to see whether candidates' performances on videos differed markedly from the classroom performances we observed. (We did not have access to both videos for all candidates, nor did we review any of the written artifacts submitted with the videos.)

Our reviewer generally agreed with the scores assigned to the videos for three of the four candidates who were recommended for certification based on their exercise scores but whose observations were not as highly scored. The videos of Candidates H, N, and O may have captured their "best practice" while their observations (for the reasons explained in the above section) may have measured performance under less than ideal teaching situations.

For those two candidates who scored better on observation than on exercises, the reviewer also generally agreed with the scores associated with their exercise videos.[2] Candidate B filmed a "writer's workshop" for her PTE submission, which was scored 2.25. Several of the classroom observers saw

[2] The third teacher in this category, Candidate J, had an unscorable video and we did not have an opportunity to review either of her videos.

similar workshops taught by Candidate B, generally assigning scores in the 3 range. On video, one can only observe the students writing, but in the classroom, the observer can read what the students have written and can watch the peer and student/teacher interactions about writing. Thus, we agree that the video did not show "accomplished teaching," but we believe that accomplished teaching may have been going on in that classroom on the day that the video was made. Candidate A's PRIDE exercise was scored a 2.00, and our observer agreed with that score. The video, however, may not be representative of what that teacher's classroom is usually like. Observers of her classroom repeatedly cited a loose, friendly, and engaged atmosphere, while the video revealed a nervous teacher and nervous students. The presence of the video camera may have disrupted the normal classroom environment.

Candidate D again presents the biggest problem in explaining differences between the exercise scoring and the observation scoring. We reviewed both her PRIDE and PTE videos and believe that neither of them demonstrates accomplished teaching. These exercises were scored 2.875 (PTE) and 3.75 (PRIDE), both over the general cut point of 2.75. This teacher received an average observation score of 1.38 and no scores above 2. Part of this discrepancy might have to do with written artifacts associated with her video-based exercises and part may be related to the production value of the video itself. For example, this teacher submitted a high-quality PRIDE video: The camera was operated with expertise, panning the class and focusing on the speaker. The class was teacher centered, with the teacher following an inflexible agenda comprised of closed-ended and largely unconnected questions; there was no student-to-student communication, and there was no student-initiated exploration—the class could be considered Standards-based only on the most superficial level.

The problem with Candidate D is not, however, only in her video-based exercise scores: her exercise-based scores are generally high while her observation scores are consistently low. Even if her PRIDE exercise, for example, had received a score 1.25 points lower, she would have been recommended for certification.

The review of these videos again suggests that exercises and observations measure different things.

Conclusion

Were our results based on a much larger sample, we would conclude that the certification process should include observation. We can conclude that classroom observation appears to be sufficiently reliable and sufficiently grounded in the Early Adolescence/English Language Arts Standards for National Board Certification to warrant inquiry as to why observation scores seem inconsistent with certification decisions. The EA/ELA Standards state, “Perhaps the single clearest sign of excellent teaching practice in any discipline can be found in the response of students.” Classroom observation may be a sensitive measure of important facets of student response not measured as thoroughly by other assessments. For these reasons, the Board may wish to consider further development of classroom observation, at least as a tool for identifying elements of accomplished practice that may be inadequately measured by extant assessments.

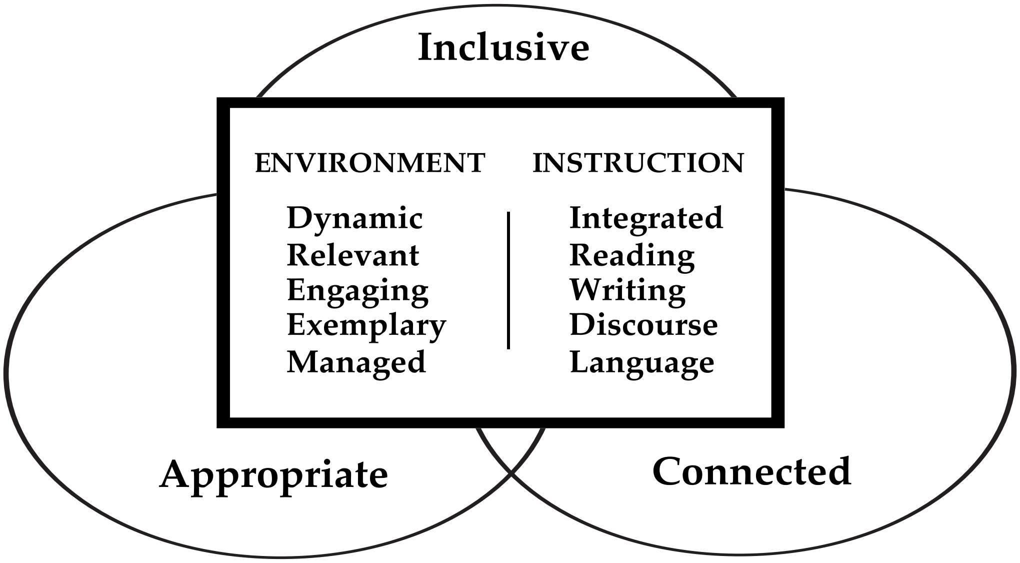
APPENDIX A

LIST OF ABBREVIATIONS

ADL	Assessment Development Laboratory
EA/ELA	Early Adolescence/English Language Arts
ETS	Educational Testing Service
NBPTS	National Board for Professional Teaching Standards
TAG	Technical Analysis Group
PRIDE	Post-Reading Interpretive Discussion Exercise
PTE	Planning and Teaching Exercise
SLE	Student Learning Exercise
IA	Instructional Analysis
CK1	Content Knowledge I: Text Selection
CK2	Content Knowledge II: Theory of Response to Literature
CK3	Content Knowledge III: Language Variation

APPENDIX B

STANDARDS CAPTURING FRAMEWORK LOGO



STANDARDS CAPTURING FRAMEWORK

APPENDIX C

NBPTS OBSERVATION INSTRUMENTS

Observer Introduction
Observation Cover Sheet
Narrative Recording Sheet
Observation Decision Sheet

CLASSROOM OBSERVATION SYSTEM:

OBSERVER INTRODUCTION

"*Precise visualization, or worse still a straining of one's attention to see crystal-clearness where there is in fact none, will only produce wrong or unusable*
results . . ." [1]

"*The moral is simple: Only partial perspective promises objective vision.*" [2]

In its policy statement, "What Teachers Should Know and Be Able to Do," the National Board for Professional Teaching Standards identifies those behaviors and practices that distinguish exemplary teachers. The Classroom Observation Study will focus on what the candidate is "Able to Do" in the classroom.

The goal for this phase of the Classroom Observation Study is to provide observers with the means to make consistent, Standards-based assessments. By "consistent," we mean that independent observers of the same performance will give the same score for the same reasons. By "Standards-based" we mean that the reasons observers cite for a particular score can be traced to language in the National Board's Standards for Early Adolescence/English Language Arts teachers.

The key to observer consistency may be to avoid fixing attention too long on any particular aspect of the teaching performance. Stated positively, the observer should keep shifting focus, keep recording and moving on. We suggest using the Standards-based categories of **Classroom Environment** and **Instruction** as focuses between which observers may alternate.

The key to making "Standards-based" assessments may be for observers to record classroom events from the stance of the principal values of the Standards as they relate to what the teacher is "Able to Do." Three practice values that seem to be infused nearly everywhere in the Standards are: **Inclusiveness, Connectedness,** and **Appropriateness.**

____

[1] Ehrenzweig, A. (1967). *The hidden order of art: A study in the psychology of artistic imagination.* Berkeley: University of California Press, p. 42.

[2] Haraway, D. (1991). *Simians, cyborgs, and women.* New York: Routledge, p. 190.

NBPTS OBSERVATION COVER SHEET

TEACHER NAME: ____

OBSERVER NAME: ____

DATE: ___________________ SCHOOL: _______________________________________

GRADE: __________________

CLASS START/END TIME:_______________/____________________

PRE-TEACHING INTERVIEW NOTES:

1. What are today's general objectives?

2. What is today's instructional focus?

3. How does today's lesson fit in with previous and upcoming lessons?

4. Is there anything special about this lesson or group of students you would like us to be aware of?

____

INSTRUCTIONAL FOCUS(ES) OBSERVED: R____; W____; D____; L____

COMMUNICATION ROLES OBSERVED (%): A____; B____; C____; D____
(**A** = T–>S; **B** = T<–>S; **C** = S<–>S + T<–>S; **D** = S<–>S<–>S<–>T<–>S . . .)

GROUPING PATTERN(S) OBSERVED: WC_____; SG_____; IS_____

POST TEACHING INTERVIEW NOTES:

1. Did today's lesson go as planned?

2. Do you plan any adjustments to future lessons as a result?

3. Other questions and notes. (Please record additional information on back, if necessary.)

Observation record sheet here

NBPTS OBSERVATION DECISION SHEET

I. Immediately after the observation, indicate your preliminary judgment by circling one of the following ratings:

1/2	2/3	3/4
(weak/adequate)	(adequate/accomplished)	(accomplished/exemplary)

II. Complete the Observation Cover Sheet, discuss class with teacher (if not possible, discuss before completing step III), and code your notes using the Standards-based Reference List Outline.

III. From the review of your notes (and referring to the Standards-based reference list), list or write about the salient features of the teacher’s performance that led you to the rating above. On your observation record sheets, number the places where you located salient features and enter the corresponding number in the column to the left headed “Ref.”

Ref.

IV. Indicate your final score by circling one of the following ratings:

1 (weak)   2 (adequate)   3 (accomplished)   4 (exemplary)

V. Use the back of this sheet to record notes and comments about your judging process.

APPENDIX D

STANDARDS-BASED REFERENCE LIST FOR OBSERVERS

STANDARDS-BASED REFERENCE LIST:
CATEGORY AND SUBCATEGORY OUTLINE

Note: Numbers in brackets [ ] show how often each item was cited as a "Salient Feature"

I. Environment (The teacher . . .); [934]
   A. Creates a dynamic of learning. [208]
   B. Relates learning activities to the interests and concerns of young adolescents. (relevant); [101]
   C. Engages all students. (engaging); [279]
   D. Sets a personal example through his or her own demeanor. (exemplary); [150]
   E. Efficiently manages the classroom. [196]

II. Instruction (The teacher . . .); [498]
   A. Integrates language arts in the creation/interpretation of meaningful texts. [72]
   B. Engages students in reading/responding to, interpreting/thinking deeply about literature and other texts. [161]
   C. Immerses students in art of writing. [120]
   D. Fosters thoughtful discourse, providing opportunities for many speaking/listening modes and purposes. [99]
   E. Uses language study to strengthen student sensitivity proficiency in appropriate uses of language. [46]

CROSS-REFERENCED AND ANNOTATED

STANDARDS-BASED REFERENCE LIST:

CATEGORIES, SUBCATEGORIES, AND DESCRIPTORS

Standards are Cross-Referenced to the Early Adolescence/English Language Arts Standards for Board Certification, September 1994. Citations for individual descriptors read: ("Standard Number". "Paragraph". "Line(s)". "Page"); Numbers in brackets [ ] show how often each subcategory and descriptor was cited as a "Salient Feature."

I. Learning Environment (The teacher . . .);

A. Creates a dynamic of learning [38].

1. maintains high expectations for the language development of all students (III.4.1-2.13) [15];
2. creates a trusting classroom environment, an atmosphere in which all students can develop competence without fear of failure or social stigmatization (IV.1.11-15.15) [37];
3. establishes classroom cultures of mutual trust and respect (IV.2.1-2.15) [27];
4. demonstrates that false starts and mistakes are part of the learning process (III.2.17-18.13) [5];
5. provides constant opportunities for students to engage actively in meaning making and expression (II.4.11-12.11) [19];
6. encourages self-directed learning while gauging student progress (II.1.9-10.11) [17];
7. makes mid-course corrections when an activity is seen to be falling flat (II.3.12-13.11) [26];
8. nests activities within a purposeful instructional framework (II.3.14-16.11) [24].

B. Relates learning activities to the interests and concerns of young adolescents (relevant) [38];

1. is familiar with youth culture and recognizes that students at this age have their own agenda (I.6.2-3.10/III.3.7-8.13) [1];
2. constantly makes connections between students' experiences of the world and language and literature (I.6.7-9.10) [21];
3. invites students to talk and write about themselves (I.4.9-10.9) [5];
4. recognizes the difference between grumbling and alienation (II.3.9-12.11) [1];
5. demonstrates an awareness that not all students learn in the same way (I.2.2-3.9) [8];
6. adjusts practice, as appropriate, based on student feedback (II.3.2-3.11) [14];
7. always leaves room for student initiated exploration (III.3.8-10.13) [13].

C. Engages all students. (engaging) [85];

1. maintains a student-centered classroom (I.1.1-2.9) [54];
2. gives students a sense of ownership (II.2.21.11) [20];
3. negotiates with students the pursuit of learning goals (II.2.19-20.11) [9];
4. encourages students to function as part of a learning community (IV.4.8-9.15) [15];
5. arranges frequent collaborative learning excises (IV.6.4-6.16) [27];
6. puts energy and creativity into capturing the interest of their students (III.3.4-6.13) [9];
7. tries many strategies to engage students; if one approach to stimulating curiosity doesn't work, they try another, and another, until they find a strategy that does (II.4.4-7.11) [21];
8. provides multiple ways into the learning process (III.5.9-10.13) [15];
9. provides opportunities for all students to use language in creative and non-threatening ways (III.4.8-10.13) [23];
10. realizes the threshold of success varies from student to student and provides alternate routes to the same learning destination (II.4.7-8.11/II.3.3-4.11) [1].

D. Sets a personal example through his or her own demeanor. (exemplary) [37];

1. relates to students as an ambassador from the adult world and is confident with the adult role (I.6.10-12.10/IV.2.11-12.15) [8];
2. is friendly to students (IV .2.4.15) [30];
3. is caring, fair-minded, and supportive of each student's well-being (IV.2.14-15.15) [25];
4. demonstrates enthusiasm about literature and the language process (III.2.1-2.13) [18];
5. is a co-learner (III.2.10-11.13) [13];
6. does not project self as authority figure (III.2.19-20.13) [4];
7. pitches leadership between too-rigid control and excessive looseness (II.3.4-6.11) [4];
8. ignores good natured irreverence (IV.2.12-13.15) [4];
9. never gives up on a student (II.4.3-4.11) [5];
10. values diversity of language experience, cultural background, heritage (IV.1.16-18.15) [2].

E. Efficiently manages the classroom [85].

1. operates with a sense of purpose in the classroom (II.1.1-3.11) [25];
2. establishes orderly and workable routines that maximize productivity and make efficient use of instructional time (IV.4.3-5.15) [38];
3. demonstrates that classroom discipline is largely a function of student engagement (IV.5.1-2.15) [8];
4. is skilled at preventing discipline problems from arising (IV.5.13-14.16) [11];
5. handles problems quickly and fairly (IV.5.17-18.16) [13];
6. minimizes disruptions to the learning process (IV.5.19-20.16) [11];

7. demonstrates the knowledge that genuine achievement motivates students to do their best (III.5.1-2.13) [5].

II. Instruction (The teacher . . .);

A. Integrates language arts in the creation/interpretation of meaningful texts [38].

1. intentionally designs learning activities that exploit mutually reinforcing tendencies of the language arts (X.2.18-20.27) [23];
2. (deleted);
3. regularly asks students to respond to intellectual challenges that require them to compose and interpret using all four language processes (X.2.20-23.27) [3];
4. integrates practice in a broad-gauged sense—organized around large, compelling themes and ideas (X.2.24-26.27) [6];
5. helps students understand that language competencies are acquired across the curriculum (X.2.27-29.27) [2].

B. Engages students in reading/responding to, interpreting/thinking deeply about literature and other texts [66].

1. uses a range of activities that permit students—regardless of their level—to demonstrate their comprehension, interpretation, and appreciation of texts (VI.6.1-5.20) [16];
2. uses texts from all media to support the development of analytical, interpretive, and critical thinking (VI.7.7-10.20) [5];
3. encourages a wide range of interpretations (VI.25-6.19) [10];
4. values all constructive responses to reading (VI.2.8-9.19) [6];
5. insists that interpretations be based on the best evidence available (VI.2.9-10.19) [10];
6. leads discussion back to the text in the case of disagreement (VI.2.11-12.19) [8];
7. asks open-ended questions about the text following logical train of student-initiated observations, rather than an inflexible agenda VI.3.3-5.19) [20];
8. helps students develop strategies for reading (VI) [10];
9. provides students an opportunity to explore value issues through reading VI.5.8-10.20) [8];
10. invites students to consider how print and non-print media differ from each other (VI.7.1-2.20) [2].

C. Immerses students in art of writing [34].

1. gives students the opportunity to write about issues which have meaning in their own lives (VII.3.3-4.21) [12];
2. connects student writing to peer audience (VII.6.6-8.21) [9];
3. gives students an opportunity to perceive one another as authors by allowing them to share their writing (VII.6.7-10.21) [9];
4. recognizes that writing is a social act (VII.6.1.21) [11];
5. helps students realize the impact their writing has on a reader (VII.7.5-6.22) [2];
6. uses student texts to present the conventions of language (VII.8.1-2.22) [1];

7. invites students to develop effective writing strategies of their own by analyzing their own and classmates' writing efforts (VII.9.27-29.22) [2];
8. shares with students different approaches, tools, and conventions used in different writing genres in order to assist students in communicating their ideas (VII.9.11-14.22) [2];
9. sponsors informal writing activities (VII.3.6.21) [7];
10. teaches writing as a recursive thinking process that can be approached systematically (VII.4.1-2.21/VII.5.3.21) [11];
11. keeps models of writing before their students and makes public the cognitive secrets that lie behind writing (VII.5.12-13.21/VII.5.3-5.21) [8];
12. chooses activities that highlight various aspects of the writing process, simplifying it into a series of relatively achievable mental tasks (VII.5.6-9.21) [5];
13. helps students think about how the composition might be changed to better fulfill their communicative intent (VII.7.4-5.22) [2];
14. shares with students their own strategies, frustrations, and insights in solving composition problems (VII.5.10-12.21) [0];
15. responds to student writing as a trusted adult interested in what the student has to say (VII.7.8-10.22) [3];
16. demonstrates a constructive response to texts which students can imitate in their reactions to one another's writing (VII.7.6-8.22) [0];
17. recognizes when students need either privacy or dialogue with a peer or their teacher (VII.9.24-24.22) [2].

D. Fosters thoughtful discourse, providing opportunities for many speaking/listening modes and purposes [39].

1. is a fluent and adept user of the spoken word (VIII.3.2-3.23) [10];
2. demonstrates effective oral re-retelling of stories (VIII.3.4-6.23) [2];
3. participates in the classroom conversation about literature or other texts (VIII.4.1-2.23) [2];
4. provides students with abundant opportunities to take part actively in challenging uses of speech (VIII.1.11-13.23) [5];
5. (deleted);
6. helps students directly with improving their speech by introducing new vocabulary and follow-up ideas designed to stretch and elevate the students' communicative competence (VIII.3.8-13.23) [4];
7. asks open-ended questions that genuinely seek information and place value on eliciting student opinion (VIII.4.5-7.23) [13];
8. listens carefully to what students have to say (VIII.4.8-9.23) [13];
9. works toward effective classroom discussion by systematically coaching it and demonstrating it, gradually giving way to independent student interaction (VIII.4.9-12.23) [10];
10. makes students aware that speech varies in different social and cultural contexts (VIII.2.6-8.23) [1];
11. uses variations in language style within the classroom community as a resource for students to learn about and appreciate language diversity (VIII.2.8-11.23) [0].

E. Uses language study to strengthen student sensitivity to/proficiency in appropriate uses of language [21].

1. employs the accepted use of grammar, syntax, and usage in their daily classroom conversations (IX.2.1-3.25) [1];
2. models and teaches conventions of English as a way of expanding each students' opportunity to participate fully in society (IX.2.18-21.25) [2];
3. incorporates activities which help students assess the different situations they find themselves in and employ English usage required by the situation (IX.4.6-8.25-26) [0];
4. helps students improve their language skills by adding to their students' communicative competencies (IX.3.2-3.25) [4];
5. celebrates the diversity of language forms (IX.1.8-9.25) [0];
6. does not try to eradicate dialectical variation from their classroom (IX.3.1-2.25) [0];
7. attends to "standard" conventions in written, rather than spoken forms (IX.3.4-5.25) [0];
8. is sensitive to students whose language reflects a non-dominant dialect (IX.2.14-17.25) [0];
9. respects the value and integrity of their students' home language (IX.2.17-18.25) [0];
10. takes care to avoid embarrassing students who are acquiring a new language (IX.6.12-14.26) [2];
11. adjusts practice to make curriculum available to ESL students (IX.5.3-5.26) [4];
12. eliminates difficult jargon from their speaking (IX.5.5-6.26) [0];
13. restates key points (IX.5.6-7.26) [1];
14. uses a slower, but natural speech rate with clear enunciation and simplified vocabulary (IX.5.7-8.26) [3];
15. accompanies explanations with pictures, objects, visual clues (IX.5.9-10.26) [1];
16. carries out regular comprehension checks (IX.5.10-11.26) [2];
17. provides needed background in pre-reading activities (IX.5.15-16.26) [2];
18. uses small groups to create "safe havens" for ESL students to gain confidence (IX.6.14-17.26) [3].

APPENDIX E

LETTER TO PARTICIPATING TEACHERS

Dear Teacher,

Under contract with the National Board for Professional Teaching Standards, the Center for the Study of Evaluation at UCLA is observing teachers who submitted EA/ELA Portfolios last year to participate in a follow-up study involving classroom observation. Participation will involve allowing pairs of trained observers to observe your teaching of two different groups of students (where possible) on three different days during October and early November. Each visit would be scheduled in advance to enable you to show how you engage your students in different types of lessons. Scheduling will accommodate your instructional plans.

This study is for research purposes to provide very important information to the Board about the costs and feasibility of including observation in its assessment packages. Observation results will be held strictly confidential and will in no way impact your candidacy for Board certification. You will receive a modest honorarium of $100.00 for your participation. We hope that you will also receive the satisfaction of knowing that you have helped forward the ambitious and worthwhile goals of the National Board for Professional Teaching Standards and made an additional contribution to the profession.

If you have questions, please feel free to contact Dr. Robert Land, Observation Project Director, at (310) 206-1532; by mail at CSE, 10880 Wilshire Blvd., Suite 700, Los Angeles, CA 90024; or by E-mail at land@cse.ucla.edu.

Thank you for your continuing support. Your professional commitment is greatly appreciated.

Yours truly,

Robert Land, Ph.D.
Project Director

APPENDIX F

NBPTS Classroom Observation Study
TEACHER DEMOGRAPHIC INFORMATION FORM

Name: ________________________________ Social Security #: ______________________

Address: ____

____

Telephone: H (___)___________________________W (___)________________________

Fax: ___________________________E-Mail _______________________________________

School: ____

Please complete **all** of the following that apply to you:

Gender: Male____ Female____

Years English Language Arts Teaching
Experience at Levels:
(Please enter number of years)

_______K-5
_______6
_______7
_______8
_______9
_______10-12
_______Community or Jr. College
_______4-year College or University

English Language Arts Teacher
Supervision Experience at Levels:
(Please enter number of years)

_______K-4
_______5
_______6-8
_______9
_______10-12
_______Community or Jr. College
_______4-year College or University

Other Subject(s) ________________________
____
____

Years Teaching Experience at Levels:
(Please enter number of years)

_______K-4
_______5
_______6-8
_______9
_______10-12
_______Community or Jr. College
_______4-year College or University

Current Roles (check any that apply):

_______Parent
_______Graduate Student
_______Teacher
_______Researcher
_______Writer
_______Teacher Educator

Other Profession:_______________________
____
____

APPENDIX G

NBPTS Classroom Observation Study
CLASS INFORMATION FORM

TEACHER NAME: ____

SCHOOL: ____

CLASS 1 START/END TIME (S): _____________/______________

CLASS 1 COURSE TITLE: ____

* CLASS 1 ABILITY LEVEL: A ______; B_______; C_____; D ____

NUMBER OF STUDENTS ENROLLED: ____________________;

MALE/FEMALE STUDENT PROPORTION: _______________.

If we are observing two different groups of students, please complete the following information for the second group.

CLASS 2 START/END TIME (S): _____________/______________

CLASS 2 COURSE TITLE: ____

* CLASS 2 ABILITY LEVEL: A ______; B_______; C_____; D ____

NUMBER OF STUDENTS ENROLLED: ____________________;

MALE/FEMALE STUDENT PROPORTION: _________.

* Compared to the overall group of students you usually teach, roughly, what percent of the students in this class are: A (High); B (High/Average); C (Low/Average); D (Low)?

____
Please use this space to record notes and comments.

APPENDIX H

NBPTS Classroom Observation Study
TEACHER TELEPHONE SURVEY FORM

TEACHER ID: ___________________________

INTERVIEWER NAME: ____

DATE: _________________

Is anything significantly different about the classes we observed compared with the classes represented in your NBPTS portfolio?

Did the presence of observers in the classroom affect the way you taught? Did it affect your students?

Did you feel that you performed about the same over the three days? If not, do you remember a particular day that seemed better or worse?

Do you think that there are days of the week that you would score better in a Classroom Observation? If so, which days?

Do you think that the three observation days give an accurate portrayal of your teaching (are they a representative sample)?

NBPTS Classroom Observation Study
TEACHER TELEPHONE SURVEY FORM (continued)

How many times do you think a teacher would need to be observed to make a valid certification decision? Why?

To what extent did Classroom Observation provide you an opportunity to demonstrate important aspects of your ability as an Early Adolescence English Language Arts Teacher?
(Not at all/To a limited extent?/Moderate extent? Substantial extent?)

What can observation do that the other exercises cannot?

What can observation NOT do? (If you were being evaluated under an observation system, what kinds of supplemental material or information would you like the observers to have?)

If observations were to be used, should teachers be required to have a passing score every day in order to be certified?

Is it relatively easy or relatively difficult to demonstrate Standards-based teaching in your particular school environment and with your particular student population?

REFERENCES

Abedi, J., & Baker, E. (1994). *A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric.* Los Angeles: University of California, Center for the Study of Evaluation/CRESST.

Baker, E., Gearhart, M., & Herman, J. (1993). *The Apple classrooms of tomorrow* (CSE Tech. Rep. No. 353). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Barrett, J. (1987). *The evaluation of teachers.* ERIC Clearinghouse on Teacher Education, ERIC Digest 12.

Berry, B., & Ginsberg, R. (1988). Legitimizing subjectivity: Meritorious performance and the professionalization of teacher and principal evaluation. *Journal of Personnel Evaluation in Education*, *2,* 123-140.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Burton, E., & Linn. R. (1994). *Comparability across assessments: Lessons from the use of moderation procedures in England* (CSE Tech. Rep. 369). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Calfee, R. C., Dunlop, K., & Watt, A. (1994). Authentic discussion of texts in middle grade schooling: An analytic-narrative approach. *Journal of Reading, 37*(7), 546-556.

Calfee, R. C., & Perfumo, P. (1993). *Student portfolios and teacher logs: Blueprint for a revolution in assessment* (Tech. Rep. No. 65). Berkeley: University of California/Carnegie Mellon University.

Chenoweth, T. (1991). Evaluating exemplary teaching. *Journal of Personnel Evaluation in Education, 4,* 359-366.

Constable, E. (1984, April 24-26). *Inter-judge reliability: Is complete agreement among judges the ideal?* Paper presented at the annual meeting of the National Council on Measurement in Education.

Crocker, L. (1992). *Matching the standards, exercises and scoring guides for the early adolescent English arts assessment by the standards committee.* Detroit: National Board for Professional Teaching Standards.

Cruickshank, D. R., & Haefele, D. L. (1990). Research-based indicators: Is the glass half-full or half-empty? *Journal of Personnel Evaluation in Education, 4,* 33-39.

Daniel, L. G., & Siders, J. A. (1994). Validation of teacher assessment instruments: A confirmatory factor analytic approach. *Journal of Personnel Evaluation in Education, 1,* 29-40.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53,* 285-328.

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge: Performance assessment. *Educational Researcher, June-July,* 11-18.

Dwyer, C. A. (1994). Criteria for performance-based teacher assessments: Validity, standards, and issues. *Journal of Personnel Evaluation in Education, 8,* 135-150.

Ellett, C. D., Loup, K. S., Evans, R. L., Chauvin, S. W., & Naik, N. S. (1994). Issues in the application of a conjunctive/compensatory standards setting model to a criterion-referenced, classroom-based teacher certification assessment system: A state case study. *Journal of Personnel Evaluation in Education, 8,* 349-375.

Ellett, C. D., Loup, K. S., Naik, N. S., Chauvin, S. W., & Claudet, J. G. (1994). A study of teachers' nominations of superior colleagues: Implications for teacher evaluation programs and the construct validity of classroom-based assessments of teaching and learning. *Journal of Personnel Evaluation in Education, 1,* 7-28.

Engelhard, G., Jr. (1994). *Maintaining the ongoing psychometric quality of the NBPTS assessments of accomplished teachers.* Commissioned paper submitted to Richard M. Jaeger, National Board for Professional Teaching Standards.

Evertson, C. M., & Burry, J. A. (1989). Capturing context: The observation system as lens for assessment. *Journal of Personnel Evaluation in Education, 2,* 297-320.

Frederickson, J. R., Sipusic, M., Gamoran, M., Wofe, E. (1992). *Video portfolio assessment. A study for the National Board for Professional Teaching Standards.* Berkeley, CA: Educational Testing Service.

Gearhart, M., Herman, J. L., Baker, E. L., Novak, J., & Whittaker, A. (1992). *A new mirror for the classroom: A technology-based tool for documenting the impact of technology on instruction* (CSE Tech. Rep. No. 336). Los Angeles: University of California, Center for the Study of Evaluation/Center for Technology Assessment.

Gordon, B., and Millman, J. (1993). *Scoring and calibration issues with a focus on reducing variability*. Paper presented to the Technical Analysis Group of the National Board for Professional Teaching Standards.

Grover, B. W. (1991). The teacher assessment dilemma: What is versus what ought to be! *Journal of Personnel Evaluation in Education, 5,* 103-119.

Haefele, D. M. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7,* 21-31.

Hancock, G. R., Shannon, D. M., & Trentham, L. L. (1993). Student and teacher gender in ratings of university faculty: Results from five colleges of study. *Journal of Personnel Evaluation in Education, 6,* 235-248.

Jacobson, L. S., & Pecheone, R. L. (1991). Connecticut teacher assessment center (CONNTAC) program: Assessing professional knowledge of beginning teachers. *Journal of Personnel Evaluation in Education, 5,* 205-226.

Jaeger, R. M. (1993). *Live vs. Memorex: Psychometric and practical issues in the collection of data on teachers' performances in the classroom*. Greensboro: University of North Carolina, Center for Educational Research and Evaluation.

Jaeger, R. M. (1993). *Setting performance standards for National Board certification. A two-stage judgmental policy capturing procedure: Some information on method and some preliminary results*. Greensboro: University of North Carolina, Center for Educational Research and Evaluation.

Jaeger, R. M. (1994). *Selected results from a pilot test of the early adolescence/ English language arts revised scoring system*. Greensboro: University of North Carolina, National Board for Professional Teaching Standards, Technical Analysis Group.

Jaeger, R. M. (1994, June 30-July 3). *Setting standards for complex performances: An iterative, judgmental policy capturing strategy*. Paper presented at the annual meeting of the American Psychological Society.

Jaeger, R. M., & Bond, L. (1993). *Issues in scoring and setting performance Standards for certification of teachers by the National Board for Professional Teaching Standards*. Greensboro: University of North Carolina, National Board for Professional Teaching Standards, Technical Analysis Group.

Jaeger, R. M., & Thompson, M. (1994). *Evaluation of the National Board for Professional Teaching Standards 1993-94 field tests of the early adolescence generalist and early adolescence English language arts assessments: A summary of responses of candidates for National Board Certification to questions posed during face-to-face interviews, focus* *group discussions, and hour-long evaluative writing tasks*. Greensboro: University of North Carolina, National Board for Professional Teaching Standards, Technical Analysis Group.

Jaeger, R. M., Bond, L., Gomey, B., & Johnson, R. (1993). *An exploratory analysis of data collected in a small-scale field test of the National Board for Professional Teaching Standards' Early Adolescence/English Language Arts assessment package*. Greensboro: University of North Carolina, Center for Educational Research and Evaluation.

Kishor, N. (1992). Compensatory and non compensatory information integration and halo error in performance rating judgments. *Journal of Personnel Evaluation in Education, 5,* 257-268.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April 5). *The effects of high-stakes testing: Preliminary findings about generalization across conventional tests*. Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Kuligowski, B., Holdzkom, D., & French, R. L. (1993). Teacher Performance evaluation in the southeastern states: Forms and functions. *Journal of Personnel Evaluation in Education, 6,* 335-358.

Linn, R. L. (1991, April). *Examinations, validity, comparability, and consequences*. Presentation at the annual meeting of the American Educational Research Association.

Linn, R. L. (1994, April). *Performance assessment: Policy promises and technical measurement standards*. Division D invited address at the annual meeting of the American Educational Research Association.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Marshall, J., Bouljon, G., Graham, P., Hynds, S., Smagorinsky, P., Smith, J., Stires, S. (1994). *A comparative analysis of the ETS and ADL scoring strategies for the National Board for Professional Teaching Standards EA/ELA assessment package*. Greensboro: University of North Carolina at Greensboro, National Board for Professional Teaching Standards, Technical Analysis Group.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observation schemes: Where are the errors? *American Educational Research Journal, 9,* 13-27.

Medley, D. M., Coker, H., Coker, J. G., Loventz, J. L., Soar, R. S., & Spaulding, R. L. (1981). Assessing teacher performance from observed competency indicators defined by classroom teachers. *Journal of Education Research*, *74*, 197-216.

Micceri, T. (1986). *Assessing the stability of the Florida performance measurement system summative observation instrument: A field study.* Unpublished paper.

Moss, P. A. (in press). Can there be validity without reliability? *Educational Researcher.*

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

National Board for Professional Teaching Standards. (1990). *Annual Report.* Detroit: National Board for Professional Teaching Standards.

National Board for Professional Teaching Standards. (1991). *Towards high and rigorous standards for the teaching profession.* Detroit: National Board for Professional Teaching Standards.

National Board for Professional Teaching Standards. (1993, June). *Rethinking assessment development: Background paper.* Paper presented at the National Board for Professional Teaching Standards Board meeting.

National Board for Professional Teaching Standards. (1993, June). *Rethinking NBPTS strategy on standards development.* Paper presented at the National Board for Professional Teaching Standards Board meeting.

National Board for Professional Teaching Standards. (1994, August). *Evaluation of the 1993-94 NBPTS field tests of the early adolescent generalist and early adolescent English language arts assessment packages.* Paper presented at the Technical Analysis Group Measurement Research Forum IV, Greensboro, North Carolina .

National Board for Professional Teaching Standards. (1994, September). *Early Adolescence/English Language Arts: Standards for National Board Certification.* Detroit: National Board for Professional Teaching Standards.

Novak, J. (1994). *Generalizability analyses.* Unpublished paper.

Pecheone, R. L., & Carey, N. B. (1990). The validity of performance assessments for teacher licensure: Connecticut’s ongoing research. *Journal of Personnel Evaluation in Education, 3,* 115-142.

Peterson, D., Kromrey, J., Lewis, A., Borg, J. (1992). Clinical pedagogy: Defining and measuring the teaching of essential and higher order thinking skills. *Journal of Personnel Evaluation in Education, 6,* 57-70.

Scriven, M. (1990). Can research-based teacher evaluation be saved? *Journal of Personnel Evaluation in Education, 4,* 19-32.

Shannon, D. M., Medley, D. M., & Hays, L. (1993). Assessing teachers' functional professional knowledge. *Journal of Personnel Evaluation in Education, 7,* 7-20.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching, In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York: Macmillan.

Shechtman, Z. (1992). Interrater reliability of a single group assessment procedure administered in several educational settings. *Journal of Personnel Evaluation in Education, 6,* 31-39.

Shulman, L. S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3-36). New York: Macmillan.

Smith, B. O., Peterson, D., & Micceri, T. (1987). *Florida performance measurement system predictive validity report.* Unpublished paper.

Soar, R. S., Medley, D. M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. *Phi Delta Kappan, 65,* 239-246.

Sulsky, L. M., & Balzar, W. (1988). Meaning and measurement of performance rating accuracy: some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497-506.

Sweeney, J. (1994). New paradigms in teacher evaluation: The SBESD model. *Journal of Personnel Evaluation in Education, 8,* 223-237.

Tyson, L., & Silverman, S. (1994). A detailed analysis of statewide teacher appraisal scores. *Journal of Personnel Evaluation in Education, 8,* 377-400.

Walberg, H. J. (1986) Synthesis of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214-229). New York: Macmillan.

Wilson, S., & Wineburg, S. S. (1988). Models of wisdom in the teaching of history. *Phi Delta Kappan, 70*(1), 50-58.

Wilson, S., & Wineburg, S. S. (in press). *Wrinkles in time and place: Using performance assessments to understand the knowledge of history teachers.* American Educational Research Journal.