Journal of Personnel Evaluation in Education 14:2 135–153, 2000

Student Surveys for School Teacher Evaluation

KENNETH D. PETERSON
Portland State University

CHRISTINE WAHLQUIST
Davis School District, Farmington, UT

KATHIE BONE
Davis School District, Farmington, UT

Abstract

Student's reports have the potential to add much useful information about school teacher quality. However, much research has centered on post secondary settings and many questions remain about the use of pupil surveys for K–12 teachers. The Davis County School District (Utah) uses pupil surveys as one teacher-chosen data source for teacher evaluation. The surveys of 9,765 students were analyzed for patterns of response. Item analysis suggests that pupils responded with reliability and validity. Some individual items are more defensible than others for conceptual and empirical reasons.

Students are central to the work of teachers, and they see teacher merit and worth from a point of view unlike those of administrators, other teachers, parents, or researchers. However, while pupil views are potentially an important consideration in school teacher evaluation, not much effort has gone into the development of principles and practices of this data source about teacher quality at the K–12 level. For example, many current authoritative guides to teacher evaluation practice (e.g., Millman & Darling-Hammond, 1990; Shinkfield & Stufflebeam, 1995) do not specify methods for including students. Very few school districts use systematic student input in their teacher assessment systems (Loup, Garland, Ellett, & Rugutt, 1996). While student perspectives are a little-used and controversial data source for school teacher evaluation, a strong case can be made to study and refine procedures for their use (Aleamoni, 1999; Peterson, 2000, www.teacher-evaluation.net). This especially is true in light of critiques that document the problems with current teacher evaluation practice that relies entirely on administrator reports (Stronge & Ostrander, 1997).

The purpose of this study was to refine a procedure to involve student views in the evaluation of public school teachers. The goal was to test survey instruments on a large scale, empirically examine survey items, determine norms to interpret future use of student views, and assess participant satisfaction with the strategies. This study intended to add to the estimates of validity and reliability of a school district teacher evaluation system.

Background

Problems with Current Teacher Evaluation Practice

The most common current practice in school district teacher evaluation systems is to include student relations as a checklist item or category for administrator reports. For example, the *Jordan Performance Appraisal System* (Jordan School District, 1995) includes a principal observation system with checklist items like "explains academic concepts," "illustrates relationships," and "encourages reluctant students." However, the use of these administrator reports, in lieu of asking the students themselves in surveys or focus groups, has been criticized as less accurate and more subject to administrator bias and halo effect ratings. Centra (1975) showed that adults are very poor raters of even college-age student views, let alone those of children and adolescents. Peterson (2000) summarized the low reliability of administrator reports, which is the hallmark of eighty years of research on the topic. In addition to technical inference and reporting problems, the social-context conflicts of this reporting role of administrators are notorious and numerous (Johnson, 1990; Lortie, 1975; Waller, 1932). For example, principals are expected to be both instructional leaders and summative judges, when the latter role clearly inhibits their ability to do the former (Lortie, 1975). Popham (1988) and Hunter (1988) debated whether a school principal can serve as an evaluator of teachers as employees (summative evaluation) and at the same time fulfill the goals of professional growth and supervision (formative evaluation). In addition, Ellett (1987) made the argument that principals charged with making summative evaluation decisions about teachers are morally responsible for providing professional growth and assistance to teachers (including client information feedback) when teachers are judged to be below standard. Altogether, these technical and sociological problems are serious enough to threaten the quality of practice in accord with the Personnel Evaluation Standards (Joint Committee, 1988) of Validity (A–4) and Reliability (A–5).

Empirical analysis of teacher evaluation designs and procedures is important. For example, Cook and Richards (1972) used multiple regression analysis to determine the contributors to ratings of teachers when the role of the rater varied—in their study, principal or university supervisor. The authors found that 29 per cent of the variance was explained by the range of teacher performances observed and reported, but 61 per cent (!) was explained by the *role* of the observer. Advocates of teacher evaluation practices (such as clinical supervision, the improvement of practice by evaluation) have a responsibility to show the empirical support for their contentions. The Cook and Richards finding, and others like it, leads to skepticism about the validity of teacher evaluation based solely upon the reports of principals (Scriven, 1981; Stronge, Helm, & Tucker, 1995). Stronge and Ostrander (1997) concluded that "classroom observation does not equal good evaluation" (p. 131) and that student ratings should be included as an optional data source for some purposes of teacher evaluation.

Student Reports of Teacher Performance

Most researchers place the systematic use of student views in teacher evaluation at under 5 per cent of school districts (Educational Research Service, 1988; Loup et al., 1996). However, many writers have argued for the importance of including student perspectives. For example, Aleamoni (1981, 1987, 1999) recommended student reports because students are a main source of information about (1) accomplishment of major educational goals, such as increased motivation, (2) rapport with the teacher, (3) elements of a classroom, such as the textbook, the homework, and methods of instruction, (4) communication between students and the instructor, and (5) consumer data for students who are able to choose their instructors. Aleamoni recognized that "Most of the research and use of the student rating forms has occurred at the college and university level" (1981, p. 110). Peterson (2000) added that students are good sources of information because they (1) know their own personal situation well, (2) have closely and recently observed a number of teachers, (3) uniquely know how pupils think and feel, (4) directly benefit from good teaching, (5) report in numbers that foster high reliability (in the 0.80–0.90 range), (6) furnish relatively inexpensive and unobtrusive information, and (7) are stakeholders and consumers of good teaching.

McGreal (1983) gave useful advice about constructing student report forms. He counseled against weaker items such as teacher "knows subject matter" and "has favorites" in distinction to better items such as "I get help when I need it" and "I feel my ideas are important." McGreal recommended in favor of student reports for formative evaluation, particularly for beginners, but against required student surveys for summative evaluation of veteran teachers. Cangelosi (1991) provided alternative formats for teachers to gather information from their pupils but said that student reports must be cost effective and not threatening for either students or teachers.

In a study of primary-grade student ratings of teachers, Haak, Kleiber, and Peck (1972) found that even students of this early age had a factor structure to their reports. The authors pointed out that pupils in their sample distinguished between liking a teacher and acknowledging her role as one who taught them something. Peterson, Stevens, and Driscoll (1990) analyzed 1,023 K–2 student surveys from 43 classrooms in five schools. Their form was a three-point-scale, eleven-item, individual colored-page book with face symbols and text administered verbally by a researcher unfamiliar to the students. Factor analysis identified factors of "learning new things," "time and support," and "ability to work in class." The global item ("this is a good teacher") well represented other items, factor scores, and total score ($r = 0.73$). Internal reliability (alpha) was 0.64. The researchers concluded that "student rating forms ... presented sufficient variance in results to suggest that primary students do discriminate among teachers' performances" (p. 171) and that "pupil reports may help to provide teachers with credible information about their impact in the classroom, an important resource not presently available" (p. 172). Peterson (1987) found that student reports had low correlations with other report sources: e.g., administrators ($r = 0.01$), teacher test scores ($r = -0.21$), and professional activity ($r = -0.01$); thus, student reports deserve a unique place in judging teacher performance.

Peterson and Stevens (1988) reported on the use of student survey reports as optional evidence for 373 K–12 teachers in a career-ladder promotion program. This technique was well accepted by teachers; more than 80 per cent elected student reports as an optional data source. In turn, the pupils rated their teachers highly. The averages on the global item (''my teacher is a good teacher'') had a mean of 4.57 on a five-point scale (SD = 0.42). The authors found that elementary grade students rated teachers higher than did older students (with statistical significance). Statistical analyses (descriptive, variance, factor, regression, reliability) showed that the global item well represented all other items, and total and factor ratings. The internal reliability (alpha) of the twelve-item, grades 7–12 instrument was 0.85. The mean correlation of teacher reports over two years was $r = 0.67$. The authors concluded that ''both the levels of usage and discrimination suggest that student reports present an important additional data source for school teacher evaluation'' (p. 29).

Ostrander (1995) concluded that '' ... the fairest and most comprehensive performance appraisals may involve multiple judges, each offering a unique perspective on teacher effectiveness'' (Abstract). In her studies, she found that students were somewhat more critical in their reports of teacher performance than were either administrators or parents. Stronge and Ostrander (1997) argued for including student reports as an optional data source because of the inadequacies of administrator reports. They recommended use of pupil report data for formative purposes and presented two model surveys (K–3 and 9–12) for practitioners and researchers.

The logistics of including student views are considerable (McGreal, 1983). Time and dollar costs for complex data require deliberation. Peterson (1989, 2000) reported that the expense of student surveys was in the $8.50 range for teachers who used them but averaged out over all teachers in the district to $4.35. Time costs to teachers are crucial, since time perhaps is teachers' most important commodity (Lortie, 1975). Data gathering also requires time and regularized procedures. Forms must be created, duplicated, distributed, and recovered. Issues of privacy, accountability, coding, security, and scheduling require much planning and expense. Scoring of student surveys involves choices of hand scoring or machine sense approaches. Also, the choice of *who* is to score the forms is important: individual teachers cannot score their own because of conflict of interest, professionals educators have credibility—but are expensive, and clerks must be managed and monitored. Record keeping also must balance confidentiality, credibility, and costs.

The politics and social context of using student views are ultimate determinants of practice. Internal politics concern debate among educators about the validity of asking students for their views. Certainly some teachers are in a better position for high ratings than are others. Also, some good teachers use student relations as part of their good practice, while other good teachers work more independently of students. External politics concern advocates for students' influence in decision making in education. Social context plays an important role when some teachers derive power from student advocates, while the status of other teachers is lessened by the introduction of student views into the teacher evaluation system.

Davis School District (Utah) Educator Assessment System

The school district that was the focus of this study (Davis School District, Utah) employs an Educator Assessment System (EAS) that requires that teachers each year furnish data about their performance to their building administrators, who then complete teacher evaluation reports. In addition, a more extensive review—including at least four different data sources—is required for (1) veteran teachers every four years, (2) beginning teachers each year, and (3) teachers on "performance assistance" remedial status. From a menu of data sources, teachers select the information that, in their judgment, best makes their case for value, merit, impact, and quality. Data sources in addition to student surveys include parent surveys, teacher tests, pupil achievement data, documentation of professional activity conceptually linked to performance, peer review of portfolios, action research projects, school improvement involvement, and information unique to an individual teacher. No claims are made that any single data source works well for, or should be required of, each teacher in the District. Teachers are assisted in their data collection by District EAS staff, i.e., information like student surveys takes teachers no more than five or ten minutes to consider, schedule, and inspect results. The purposes of this complex system are (1) to have teachers more involved in their own evaluation (Peterson & Chenoweth, 1992), (2) to base teacher evaluation on the best objective evidence available, (3) to gather information about performance to use in public relations, staff development, and dissemination of best practices, (4) to replace time-consuming—and not respected (Wolf, 1973)—annual formal classroom evaluation visits by the principal, and (5) to include legitimate stakeholders in the teacher evaluation system (Mark & Shotland, 1985).

The Davis School District EAS design provides multiple and variable data sources for teacher evaluation. Each teacher must present a constellation of information showing that his or her performance is well functioning. The varied configuration of data types allows for documentation of good teaching of different types and in different settings. It also largely solves the political (McGreal, 1983) and social-context (Lortie, 1975) problem of teacher acceptance of data sources. Detractors of any individual data source (e.g., in the case of student surveys, those who hold that "They are a popularity contest" or "Ratings can be easily manipulated") are not forced to take a stand that conflicts with the more numerous teacher supporters of student data. This teacher selection avoids the problem of "all or nothing" data source selection that for most school districts results in gathering student views for none of the teachers. An additional control feature for teachers in the EAS system is that teachers get two levels of control: they first must elect to have the surveys collected, and second they must elect to present the results *after* inspecting them.

The Study

Questions Addressed in the Study

This study of student surveys for teacher evaluation addressed six questions significant to the development of valid and reliable use of this data source. These issues included

selection of survey items, use of a single ''global item,'' methods of reporting scores to teachers and administrators, selection of logistics, concerns about validity and reliability, and participant satisfaction with procedures.

Item selection is an important issue for survey development. While many participants want to know about many different topics from students, not all items work equally well. Peterson (2000) described the variety of tests that possible items must pass before they are acceptable or valid for use with students. These tests include logic, reason, fairness, and finally, empirical results. It is important that students be asked to report things that they have directly experienced, for example, that the teacher provides them an opportunity to learn new things. It is equally important that students *not* be asked to report on issues they have not experienced, for example, that the teacher is equally fair to all other students or knows the subject matter well. Fairness tests of items are important because teachers vary in their performance conditions, e.g., some teachers deal with 25 children per day while others face 125. This study included a goal of empirically testing a variety of items for use in a final, most defensible form.

A second issue to be studied in this use of student surveys was the strategy of using a ''global item'' (Peterson et al., 1984). While surveys present an opportunity to question students about a wide range of topics, this variety raises problems in summative uses of the information. One way to understand student views is to report the summary statistics (mean, distributions) for each item. However, not all items are equal in value, for example, ''Rules in class help me to learn'' may or may not be as important as ''I learn new things in class.'' Alternate reports to listing each item are total scale score, average score of each item, factor scores, weighted item scale score, and global item. The problem with these alternatives is that each presents an inherent value, with advantages and disadvantages, in balancing the contribution of individual items. A single global item, such as ''this is a good teacher'' or ''rate overall performance'' can be a useful summative report, if the item can be shown to well represent the other items. The empirical interest in this study was the behavior of the global item, i.e., does it defensibly represent other items, underlying factor structure, and internal reliability of the scale.

How to report student survey scores is a crucial question for the success of using this data source. Summary statistics (mean, standard deviation, numbers of responses) on each item give absolute information about student responses. However, the perspective of comparison with other teachers is needed. Thus, district norms for item response and percent return rate are needed for each level. A specific problem is teacher rejection of scores that are not perfect (Peterson, 2000). Thus, a categorical report is recommended (e.g., ''well functioning'' or ''not well functioning''). Categories call for a cutoff score. Possibilities for a cutoff score are some absolute number, a score that would define an absolute per cent, a scree test that would eliminate laggards, or a population parameter such as a standard deviation below mean. This study examined the latter strategy for satisfactory discrimination.

Utah state law requires that a teacher evaluation system be ''valid and reliable'' (Utah State Code, 1953). This study examined the proposition that student views are a valid and reliable component of the larger teacher evaluation system. The argument that the entire evaluation system is valid and reliable is partly based on the argument that each of the

components is valid and reliable for itself. While school districts are generally not held to standards of perfection in their teacher evaluation systems (Chance v. Bd. of Examiners, 1971; McCoy, 1998), it is a reasonable expectation that they strive to use state-of-the-art techniques, test their assertions, document their actual practice, and question participants about their satisfaction.

Methods

This study used surveys with scale items to assess the views of students concerning the performance of their teachers. Item descriptors for the surveys are presented in tables 1–3 below (sections 1.1, 2.1, and 3.1, respectively).

Sample. Student surveys were elected by 401 teachers (57 K–2, 89 elementary, 255 middle- and high-school) from 27 schools in a Utah school district pilot study. Item and age-level analyzes was performed on a total of 9,765 useable surveys.

Survey data. Data for this study consisted of independent variables of teacher, level taught, and school. The dependent variable of item was a score on a three-point scale. Data were collected on bubble-mark, computer-sensed forms. Forms for primary grade nonreaders featured a cartoon-face coding scheme.

Survey analysis. Analysis began with each survey item from each level being analyzed for descriptive statistics: mean, standard deviation, kurtosis, and skewness. Next, tests of suitability for factor analysis (Bartlett test of sphericity) and sampling adequacy (Kaiser–Meyer–Olkin) were performed. Analysis of each of the scale items consisted of an intercorrelation matrix and a factor analysis using a principal components analysis with first a single-factor solution and then a varimax rotation for multiple factors. The internal reliability of each scale was done by computing Cronbach's alpha. An analysis of variance of the global item was done to assess the effects of school and level of teaching, relative to the variance contributed by the individual teacher. For smaller samples of teachers, correlations between (1) two years of data and (2) student and parent surveys on the global item were computed.

Participant satisfaction data. Participant satisfaction was assessed with a random survey of a sample of 561 teachers, equally representing elementary, middle, and high school levels. These participants were surveyed about their satisfaction with the pilot teacher evaluation system, which included the opportunity to sample student views. Descriptors of the four satisfaction survey items are presented in table 7 (see below). Administrator perspectives were documented with interviews of nine principals and four focus groups having twenty-eight participants.

Findings

Descriptive Statistics of Survey Items

Tables 1, 2, and 3 present the analysis of survey items at each grade level. All distributions were skewed toward high agreement or satisfaction. Most items were leptokurtic or "bunched" near the mean in distribution.

Table 1. Item Analysis of Primary-Pupil Survey ($n = 1,065$).
1.1. Descriptive Statistics of Primary Pupil Survey Items.

Item (Descriptor)	*Mean*	*S.D.*	*Kurtosis*	*Skewness*
1. Shows me how to do new things	2.84	0.43	6.93	− 2.72
2. Rules in class help me to learn	2.78	0.48	3.79	− 2.12
3. Learn new things in class, can tell	2.81	0.46	5.53	− 2.48
4. My teacher is a good teacher	2.90	0.36	15.15	− 3.89
5. My teacher is nice to me	2.85	0.41	6.92	− 2.70
6. I know what I'm supposed to do	2.83	0.44	6.73	− 2.69
7. Class is a good place for learning	2.87	0.43	10.47	− 3.33

1.2. Interitem Correlations of Primary-Pupil Survey Items.

	I1	*I2*	*I3*	*I4*	*I5*	*I6*
Item 2	0.25	—				
Item 3	0.31	0.28	—			
Item 4	0.32	0.28	0.36	—		
Item 5	0.30	0.29	0.33	0.47	—	
Item 6	0.21	0.35	0.24	0.32	0.32	—
Item 7	0.35	0.34	0.33	0.42	0.44	0.28

1.3. Factor Analysis of Primary-Pupil Survey Items.

	Load Factor 1	*Load Factor 2*	*Load Factor 3*	*Load Factor 4*	*Corr. with Total*	*Corr. with Item 4*
Item 1	0.20	0.09	0.90	0.13	0.43	0.32
Item 2	0.05	0.80	0.29	0.16	0.46	0.28
Item 3	0.23	0.14	0.14	0.94	0.47	0.36
Item 4	0.76	0.12	0.11	0.23	0.55	—
Item 5	0.80	0.17	0.10	0.09	0.54	0.47
Item 6	0.35	0.76	− 0.09	0.04	0.43	0.32
Item 7	0.60	0.22	0.41	0.07	0.55	0.42

Table 2. Item Analysis of Elementary-Pupil Items ($n = 2,218$).
2.1. Descriptive Statistics of Elementary-Pupil Survey.

Item (Descriptor)	*Mean*	*S.D.*	*Kurtosis*	*Skewness*
1. Shows me how to do new things	2.77	0.45	1.77	− 1.68
2. Rules in class help me to learn	2.53	0.62	− .15	− 0.94
3. Learn new things in class, can tell	2.59	0.59	0.22	− 1.11
4. My teacher is a good teacher	2.82	0.45	6.16	− 2.60
5. My teacher is nice to me	2.76	0.49	3.01	− 1.95
6. I know what I'm supposed to do	2.76	0.46	1.75	− 1.66
7. Class is a good place for learning	2.71	0.52	1.42	− 1.53
8. I know how well I'm learning	2.62	0.61	0.70	− 1.34
9. Teacher treats me with care, respect	2.76	0.52	3.70	− 2.15

2.2. Inter-item Correlations of Elementary-Pupil Survey Items.

	I1	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*
Item 2	0.25	—						
Item 3	0.32	0.25	—					
Item 4	0.42	0.34	0.33	—				
Item 5	0.38	0.35	0.34	0.66	—			
Item 6	0.17	0.22	0.28	0.21	0.21	—		
Item 7	0.36	0.36	0.32	0.43	0.41	0.19	—	
Item 8	0.26	0.30	0.34	0.27	0.29	0.31	0.29	—
Item 9	0.41	0.33	0.37	0.68	0.71	0.22	0.43	0.32

2.3. Factor Analysis of Elementary-Pupil Survey Items.

	Load Factor 1	*Load Factor 2*	*Load Factor 3*	*Load Factor 4*	*Corr. with Total*	*Corr. with Item 4*
Item 1	0.25	0.85	0.09	0.10	0.49	0.42
Item 2	0.20	0.04	0.18	0.88	0.46	0.34
Item 3	0.21	0.45	0.61	− 0.03	0.49	0.33
Item 4	0.82	0.22	0.09	0.18	0.65	—
Item 5	0.86	0.13	0.14	0.16	0.65	0.66
Item 6	0.11	0.07	0.16	0.11	0.34	0.21
Item 7	0.29	0.48	0.11	0.54	0.53	0.43
Item 8	0.16	0.00	0.87	0.23	0.46	0.27
Item 9	0.85	0.19	0.18	0.13	0.67	0.68

These tables also show the intercorrelations of survey items for each of the three grade levels. Individual item correlations ranged from $r = 0.17$ (''Teacher shows me how to do new things'' and ''I know what I'm supposed to do in class'') to $r = 0.71$ (''Teacher treats me with care and respect'' and ''Teacher is nice to me''), both at the elementary level.

Table 3. Item Analysis of Secondary-Pupil Survey ($n = 6,482$).

3.1. Descriptive Statistics of Secondary Pupil Survey Items.

Item (Descriptor)	*Mean*	*S.D.*	*Kurtosis*	*Skewness*
1. Shows me how to do new things	2.63	0.58	0.71	− 1.31
2. Rules in class help me to learn	2.39	0.71	− 0.74	− 0.73
3. Learn new things in class, can tell	2.60	0.60	0.44	− 1.22
4. This teacher is a good teacher	2.72	0.53	2.33	− 1.80
5. Cares about how well I do in class	2.60	0.63	0.60	− 1.33
6. I know what I'm supposed to do	2.71	0.51	1.65	− 1.60
7. Class is a good place for learning	2.61	0.59	0.48	− 1.23
8. I know how well I'm learning	2.56	0.63	0.17	− 1.13
9. Teacher treats me with care, respect	2.64	0.60	1.07	− 1.48
10. I know why we learn in this class	2.52	0.67	− 0.12	− 1.05
11. I understand how to do assignments	2.51	0.61	− 0.25	− 0.87
12. Teacher explains so I can learn	2.64	0.59	0.89	− 1.39
13. Helps me see how things fit	2.57	0.63	0.23	− 1.16

3.2. Interitem Correlations of Secondary-Pupil Survey Items.

	I2	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*	*I12*	*I13*
I1	0.45	0.47	0.53	0.45	0.36	0.51	0.42	0.47	0.38	0.39	0.54	0.53
I2	—	0.45	0.43	0.42	0.36	0.50	0.43	0.44	0.40	0.38	0.46	0.47
I3		—	0.49	0.40	0.37	0.51	0.48	0.39	0.44	0.44	0.52	0.52
I4			—	0.54	0.40	0.59	0.44	0.60	0.39	0.45	0.61	0.59
I5				—	0.35	0.48	0.42	0.59	0.35	0.39	0.51	0.51
I6					—	0.42	0.40	0.36	0.39	0.47	0.46	0.42
I7						—	0.47	0.51	0.44	0.46	0.58	0.57
I8							—	0.42	0.42	0.45	0.46	0.49
I9								—	0.37	0.40	0.53	0.53
I10									—	0.46	0.45	0.47
I11										—	0.56	0.50
I12											—	0.64

3.3. Factor Analysis of Secondary-Pupil Survey Items.

	Load Factor 1	*Load Factor 2*	*Load Factor 3*	*Load Factor 4*	*Corr. with Total*	*Corr. with Item 4*
Item 1	0.73	0.25	0.03	0.12	0.64	0.53
Item 2	0.27	0.25	0.15	0.14	0.60	0.43
Item 3	0.66	0.03	0.36	0.07	0.64	0.49
Item 4	0.60	0.55	0.14	0.17	0.71	—
Item 5	0.22	0.79	0.18	0.12	0.63	0.54
Item 6	0.19	0.16	0.14	0.89	0.55	0.40
Item 7	0.58	0.33	0.22	0.20	0.71	0.59
Item 8	0.26	0.23	0.44	0.20	0.62	0.44
Item 9	0.27	0.79	0.15	0.13	0.66	0.60
Item 10	0.18	0.18	0.83	0.09	0.58	0.39
Item 11	0.35	0.20	0.56	0.48	0.62	0.45
Item 12	0.64	0.36	0.30	0.31	0.75	0.61
Item 13	0.60	0.38	0.34	0.17	0.74	0.59

Factor Analysis

The test for appropriateness of using factor analysis (Bartlett test of sphericity) showed values representing probabilities of less than 0.01 that there actually were not discrete factors existing in each of the three levels. Sampling adequacy (Kaiser–Meyer–Olkin) was "meritorious" for the primary (0.85) and elementary (0.88) levels and "marvelous" for the secondary (0.96) items (Kaiser, 1974).

The factor structure of each of the three grade levels is presented in sections 1.3, 2.3, and 3.3 of tables 1, 2, and 3, respectively. A summary of differences in the factor structures appears in table 4. The most notable difference in analysis of the factor structures by level is that while all levels separate "Teacher shows caring and respect" from "Student learns," the primary and elementary levels were more concerned with the former in their rating of teachers, while the secondary more emphasized the latter. For contrast, the single-factor solution of the factor analysis for the secondary survey is presented in table 5.

Table 4. Factor Structure of Pupil Survey Items by Level—Eigenvalues and Percentage of Variance Explained for Strongest Common Factors.

	Teacher Cares and Respects, Climate	*Tells, Shows How to Do, Explains*	*Pupil Learns, Teacher is Effective*	*Pupil Knows What to Do, Clarity*
Primary	2.95	0.76	0.69	0.86
	42.2%	10.8%	9.9%	12.3%
Elementary	3.84	1.09	0.79	—
	42.7%	12.1%	8.7%	
Secondary		*and Pupil*		
	0.90	*learns*	—	0.71
	6.9%	6.59		5.5%
		50.7%		

Table 5. Single-Factor Solution for Secondary-Level Survey Factor Analysis.

		Factor 1 Load
1.	Teacher shows me how to do new things	0.707
2.	The rules in class help me to learn	0.664
3.	I learn new things in this class that I can tell you about	0.702
4.	This teacher is a good teacher	0.773
5.	This teacher cares about how well I do in this class	0.696
6.	I know what I am supposed to do in this class	0.611
7.	This class is a good place for learning	0.767
8.	I know how well I'm learning in this class	0.677
9.	Teacher treats me with care and respect	0.719
10.	I know why we are learning what we learn in this class	0.638
11.	I understand how to do the assignments in this class	0.684
12.	Teacher explains things so that I can learn	0.802
13.	Teacher helps me to see how things fit together	0.792

The reliability analysis for all scale items together for each of the three levels showed scale Cronbach's alphas of (respectively) 0.766, 0.820, and 0.917. The increase in internal reliability reflected the increasing number of items used in the three scales.

Global Item

Item 4 (overall satisfaction—"my teacher is a good teacher"), relative to other items, showed (1) high average loadings on the first three factors (0.33, 0.38, and 0.43, respectively), (2) high correlations with total of all items (0.55, 0.65, and 0.71, respectively), and (3) highest average correlation with each of other items (0.36, 0.42, and 0.51, respectively).

Table 6 presents the contribution (eta statistic derived from analysis of variance) of each independent variable (school, level, teacher) on the dependent variable of the global item ("This is a good teacher"). The independent variable of teacher explained 17.9 per cent (eta squared) of the variance in the dependent variable of the global item. The percentage of variance in teacher rating explained (eta squared) a range of values from 2.0 per cent for level to 21.4 per cent for teacher.

A pilot sample ($N = 228$) was analyzed for changes in rating form scores on the global item (#4) of the student reports for all three levels. The correlation (product–moment) between the two years was $r = 0.48$. This value may be compared with the (Peterson, 1988) report of $r = 0.67$ (obtained from a five-point scale). The means and standard deviations were 2.762 (0.259) for year one and 2.800 (0.224) for year two. A statistical test of difference between years for each of the audiences showed a $t = 2.33$ and an associated probability value of 0.02. More teachers showed changes in the positive direction (118) than in the negative (96), although not with statistical significance (sign test: $z = 1.44$, $p = 0.15$).

This study included computation of a correlation between parent and student survey ratings on the global item when both were gathered in the same year by the individual teacher. In this study the product–moment correlation was found to be $r = 0.478$ ($N = 70$). This finding compares with the Peterson (1987) value of $r = 0.584$ ($N = 243$) using a five-point scale.

Table 6. Sources of Variance in Pupil Ratings of Teacher (ANOVA).

Source	*Eta*	*Eta*^{2}
School (combined)	0.145	0.021
Primary	0.153	0.023
Elementary	0.151	0.023
Secondary	0.130	0.017
Level	0.141	0.020
Teacher (combined)	0.463	0.214
Primary	0.344	0.119
Elementary	0.545	0.297
Secondary	0.499	0.249

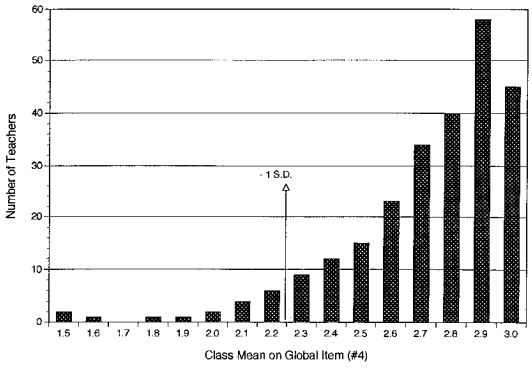

Figure 1. Secondary teacher class mean distribution and cut score location.

Cut Scores

As described earlier in this articles. Davis School District EAS teachers are required to select and present multiple data sources (www.davis.edu/staffdev/eas/index.htm). The District Evaluation Development Committee selected a number of "cut scores" on various measures (e.g., parent surveys, teacher tests) below which a data source will not qualify for consideration as one of the number required. This study investigated the implications of several levels of cut scores by the population distribution. Figure 1 presents a cut score of minus one standard deviation on the secondary student ratings. For the secondary level, a cut score of one standard deviation below the mean would exclude 6.0 per cent of the teachers from the "Well Functioning" category report; a cut score of 1.5 standard deviations below the mean would exclude 3.0 per cent. For the primary grades, a cut score of one standard deviation below the mean would exclude 4 per cent of the teachers from the "Well Functioning" category report; a cut score of 1.5 standard deviations below the mean would exclude 2 per cent.

Participant Satisfaction

Table 7 reports four items from the teacher satisfaction survey. In general, teachers were satisfied with the opportunity to gather data of their own choosing, including student

Table 7. Teacher Satisfaction with New Data System.

Item	*Agree*	*Disagree*
I had more control over my evaluation	463 (82.5%)	98 (17.5%)
Better helped me reflect on my teaching	471 (82.5%)	100 (17.5%)
Improvement over old data system	476 (84.5%)	87 (15.5%)
Yearly formal observations are not necessary for experienced, well-functioning teachers	482 (83.2%)	97 (16.8%)

views. This view is important in light of the strong sentiment that formal observations are not necessary for experienced, well-functioning teachers.

Administrators were equally positive about the new multiple data source evaluation system, but they were less sanguine about the student survey data. Principals were quick to point out potential problems, such as inadequate sampling and students not taking the process seriously. They also reported that they learned little from the surveys that they did not already know.

Discussion and Conclusions

This study supported the contention that student surveys can be valid and reliable data sources for teacher evaluation. The patterns of response, especially as disclosed by the factor analysis, suggest that students responded to the range of items with reason, intent, and consistent values. Thus, data gathered by student surveys define one dimension of teacher quality as expressed by student views, which may vary in its importance from one teacher to another. The positive response of teachers and administrators found in this study is important in light of the general dissatisfaction that teachers have with evaluation based solely on principal reports (Kauchak, Peterson, & Driscoll, 1985; Lortie, 1975; Peterson, 2000; Wolf, 1973).

The discrimination in the factor structures between a teacher as a source of learning and as a person who shows respect and care supports the validity of the data source: students at different ages distinguish between these two dimensions. Student surveys are not merely popularity contests; students distinguish between merely liking a teacher and recognizing one who enables their learning. While students can distinguish between a teacher who supports learning and one who treats them well, this study suggests that the former is more important to older students, while the latter is more important to younger ones. Another interesting finding for the high school students is that they appeared to distinguish between a teacher who explains or tells and one who fosters more student-centered learning.

One surprising set of findings led to a change in recommendations for specific items to include in a student survey. The original survey designed for this study did not include "popularity" items (e.g., "Child treated as an individual"), which may lead to ratings more based upon trivial and superficial but pleasing teacher performances, sometimes called "pandering," rather than more defensible, substantial performances, such as fostering student learning in the classroom, enabling home support of learning, and providing important information for students as legitimate stakeholders (Epstein, 1985; Peterson, 1989a; Peterson, 1995; Scriven, 1973a, 1973b, 1988). However, due to input from its multiple authors, the actual survey used in this study included some items not tested in previous studies (e.g., "Child treated as an individual"). The factor analysis of themes that underlay the literal item reports suggested an unexpectedly strong sentiment on the part of students for *caring and respectful personal treatment of pupils* (see tables 3 and 4).

The global item ("This is a good teacher") well represented the other items, scale, and factor structures. For example, the global item had the highest interitem correlations and

the highest average load on factors. This finding supports use of the global item as a single report indicator of student opinions of teachers. That is, evidence was found that the most central views of the students could be compressed by recognizing the responses to the most conceptually pertinent item (''Overall satisfaction with this teacher'') rather than a scale composite (e.g., gross total of items, weighted scale, factor score) or another single item. Each scale composite suffers in comparison to the global item because of problems such as (1) counting all items as equally important (e.g., ''Enables learning'' vs. ''Is immediately available when called upon''), (2) claims that one weighting scheme (yours) is always better than another (mine), or (3) one underlying scale factor (kind, humane treatment) is always more important than another (gives clear instruction).

The correlation levels between two years of survey usage suggest that teachers require more than two years to show stable patterns of survey results. This means that reliable estimates require the district to collect survey results for three years and then in alternate years as the patterns are established. The absolute changes in ratings could be explained by a number of possibilities, including that the teachers changed their instructional effectiveness. Other explanations include (1) teachers' experience enabled them to better prepare students and parents for more positive responses, or (2) teachers merely altered their practices to get better ratings without becoming better instructors. The finding of a relatively low correlation between student and parent surveys supports the validity of using multiple perspectives and measures of teacher quality for both formative and summative evaluation.

Recommendations

The three-point response scale was discriminating enough to make item analysis and comparisons possible to better understand the survey. However, some analyses, particularly internal reliability, showed a relatively low variability. A five-point scale is recommended for better discrimination both quantitatively and qualitatively since it permits a distinction between degrees of agreement and disagreement for those students who actually show such power of discrimination (i.e., middle- and high-school students).

This set of surveys was computer scored. Although this practice leads to increased dollar and administrative time costs, it does save logistical and record keeping costs as well as clerk time in scoring. The machine scoring of *all* items compares favorably with the hand scoring of *only* the global item by more immediately providing formative feedback information to teachers.

While the internal reliability of the survey used in this study was estimated, the important questions of consistency of recording opinion over time periods (both brief patterns—survey/resurvey—and career-long patterns—annual consistency) were not documented in this study. These are important empirical issues for the future.

The limits of student responses should be taken into account in making final judgments based upon them. For example, the amount of variance explained by teachers was high

(relative to dependent variables in educational research) but certainly not absolute: error terms and unknown sources of variance remain substantial. For this reason, and others, it is difficult to explain the reasons for low ratings. Cautions should be used: high student ratings do not necessarily mean the same thing as good teaching. Perhaps the best interpretation is that high student ratings in conjunction with at least several other positive indicators are a good indicator of quality teaching. This view is consistent with that expressed by Glass (1974), Peterson (1984), Epstein (1985), and Ostrander (1995), i.e., that multiple judges (or multiple data sources) are required to best identify teacher quality.

Student views of the teacher are important both for perceptions within a school system and as accurate indicators of performance. Our advice to teachers is to be concerned with relations to students as important people. While this can be difficult with large classes, or multiple classes with large numbers of students during the day, these characteristics are important for student judgment about teacher quality.

The need for future studies about student surveys is suggested by the findings of this study. A case study of the teachers who received low class mean scores would be helpful to better understand the meaning of low ratings by students. Also, a more discriminating factor analysis study that included a greater number of similar items for student rating would better clarify the construct of quality teaching as expressed by student surveys.

A list of recommended items appears in the Appendix.

Appendix: Recommended Items for Student Surveys

Primary Grade (K–2), prereaders:

The prereader forms should be as visual as possible. We recommend printing each item on a separate colored page, to make a 7- or 8-page booklet (about $3'' \times 8''$). The making codes can be faces: smile for "Yes," straight mouth line for "Sometimes," and frown mouth for "No."

The items can be read one by one by the clerk, who checks to make sure all students are on the same color. A large chart for the clerk to point to is helpful.

The following items have worked well:

I am able to do the work in class
Teacher is kind and friendly
I learn new things in this class
My teacher is a good teacher
Teacher shows us how to do new things
The rules in class help us to learn
I know what I am supposed to do in class
Elementary = Student Survey Form (Readers)

	Agree	Not Sure	Disagree
I know what I'm supposed to do in class	3	2	1
Teacher shows us how to do new things	3	2	1
There is enough time to finish my work	3	2	1
This class is not too noisy or rowdy for learning	3	2	1
I learn new things I can tell you about	3	2	1
I know how well I'm doing in class	3	2	1
This is a good teacher	3	2	1
Teacher treats me with care and respect	3	2	1
The rules in class help me to learn	3	2	1
These items may be added if you wish:			
We have enough materials and supplies to learn	3	2	1
This class is not too slow or fast to learn well	3	2	1

Middle- and High-School Student Survey Form

	Agree		Not Sure	Disagree	
I know what I'm supposed to do in class	5	4	3	2	1
Teacher shows us how to do new things	5	4	3	2	1
There is enough time to finish class work	5	4	3	2	1
This class is not too noisy or rowdy for learning	5	4	3	2	1
I learn new things I can tell you about	5	4	3	2	1
I know how well I'm doing in class	5	4	3	2	1
This is a good teacher	5	4	3	2	1
We have enough materials and supplies to learn	5	4	3	2	1
At the end of class, I understand well enough to finish the assignment	5	4	3	2	1
I know why we learn what we learn in class	5	4	3	2	1
This class is not too slow or fast to learn well	5	4	3	2	1
The rules in class help me to learn	5	4	3	2	1
	Agree		Not Sure	Disagree	

References

Aleamoni, L.M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110–145). Beverly Hills, CA: Sage.

Aleamoni, L.M. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education, 1*, 111–119.

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153–166.

Cangelosi, J.S. (1991). *Evaluating classroom instruction.* New York: Longman.

Centra, J.A. (1975). Colleagues as raters of classroom instruction. *Journal of Higher Education, 46,* 327–337.

Cook, M.A., & Richards, H.C. (1972). Dimensions of principal and supervisor ratings of teacher behavior. *Journal of Experimental Education, 41*(2), 11–14.

Drake, T.L. & Roe, W.H. (1994). *The principalship.* (4th ed.). New York: Macmillan College Publishing.

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards, and issues. In A.J. Shinkfield & D. Stufflebeam (eds.), *Teacher evaluation: Guide to effective practice* (pp. 62–80). Boston: Kluwer Academic Publishers.

Educational Research Service (1988). *Teacher evaluation: Practices and procedures.* Arlington, VA: Educational Research Service.

Ellett, C.D. (1987). Emerging teacher performance assessment practices: Implications for the instructional supervision role of school principals. In W. Greenfield (Ed.), *Instructional leadership: Concepts, issues, and controversies* (pp. 302–327). Boston: Allyn and Bacon.

Epstein, J.L. (1985). A question of merit: Principals' and parents' evaluations of teachers. *Educational Researcher, 14*(7), 3–8.

Glass, G.V. (1974). A review of three methods of determining teacher effectiveness. In H.J. Walberg (Ed.), *Evaluating educational performance* (pp. 11–32). Berkeley, CA: McCutchan.

Johnson, S.M. (1990). *Teachers at work: Achieving success in our schools.* New York: Basic Books.

Joint Committee on Standards for Educational Evaluation (1988). *The personnel evaluation standards: How to assess systems for evaluating educators.* Newbury Park, CA: Corwin Press.

Jordan School District (1995). *Jordan Performance Appraisal System.* Sandy, UT: Jordan School District, Utah.

Kauchak, D., Peterson, K., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19*(1), 32–37.

Lortie, D. (1975). *Schoolteacher: A sociological study.* Chicago: University of Chicago Press.

Loup, K., Garland, J., Ellett, C., & Rugutt, J. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education, 10,* 203–226.

McCoy, M.T. (1998). Reasons for teacher evaluation. Education Law Association, Spring Conference, Snowbird, Utah, March 24, p. 24.

McGreal, T.L. (1983). *Successful teacher evaluation.* Alexandria, VA: Association for Supervision and Curriculum Development.

Millman, J., & Darling-Hammond, L. (Eds.) (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* Newbury Park, CA: Sage.

Ostrander, L.P. (1995). *Multiple judges of teacher effectiveness: Comparing teacher self-assessments with the perceptions of principals, students, and parents.* Doctoral dissertation. University of Virginia, Charlottesville, VA.

Peterson, K., Gunne, M., Miller, P., & Rivera, O. (1984). Multiple audience rating form strategies for student evaluation of college teaching. *Research in Higher Education, 20,* 309–321.

Peterson, K.D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24,* 311–317.

Peterson, K.D. (1989). Costs of school teacher evaluation in a career ladder system. *Journal of Research and Development in Education, 22*(2), 30–36.

Peterson, K.D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Peterson, K.D., & Chenoweth, T. (1992). School teachers' control and involvement in their own evaluation. *Journal of Personnel Evaluation in Education, 6,* 177–189.

Peterson, K.D., & Stevens, D. (1988). Student reports for schoolteacher evaluation. *Journal of Personnel Evaluation in Education, 1,* 259–267.

Peterson, K.D., Stevens, D., & Driscoll, A. (1990). Primary grade student reports for teacher evaluation. *Journal of Personnel Evaluation in Education, 4*, 165–173.

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education, 1*, 319–334.

Shinkfield, A.J., & Stufflebeam, D. (1995). *Teacher evaluation: Guide to effective practice*. Boston: Kluwer Academic Publishers.

Stronge, J., & Ostrander, L. (1997). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 129–161). Thousand Oaks, CA: Corwin Press.

Waller, W. (1932). *The sociology of teaching*. New York: Wiley & Sons.

Wolf, R. (1973). How teachers feel toward evaluation. In E. House (Ed.), *School evaluation: The politics and process* (pp. 156–168). Berkeley, CA: McCutchan.